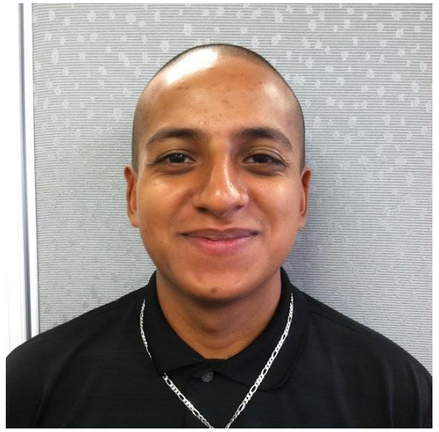
St. Jude Children's Research Hospital®
Finding cures. Saving children.

Pitfalls of sc/snRNA-seq

October 30, 2025

Sharon Freshour, PhD
Department of Developmental Neurobiology
St. Jude Children's Research Hospital

The DNB Bioinformatics Core Team



Cody Ramirez, PhD

Senior Bioinformatics Research Scientist
Core Director
Boston, Massachusetts

Antonia Chroni, PhD

Senior Bioinformatics Research Scientist
New York, New York

Asha Jacob Jannu, MS

Bioinformatics Research Scientist
Indianapolis, Indiana

Sharon Freshour, PhD

Bioinformatics Research Scientist
Tacoma, Washington

Pitfalls of sc/snRNA-seq workshop outline

- Purpose: highlight analysis steps where pitfalls occur
- Explore, fine tune, especially if data returns unexpected results
- Going to cover:
  - Cell Level QC and Filtering
  - Normalization, Variable Feature Selection, Scaling, Regression
  - Dimensionality Reduction, Clustering, Visualization
  - Integration
  - Differential Expression Analysis
  - Cell Type Annotation
- Focused on analysis in Seurat, but pitfalls are general to any sc/snRNAseq analysis

Pre-Seurat QC Steps

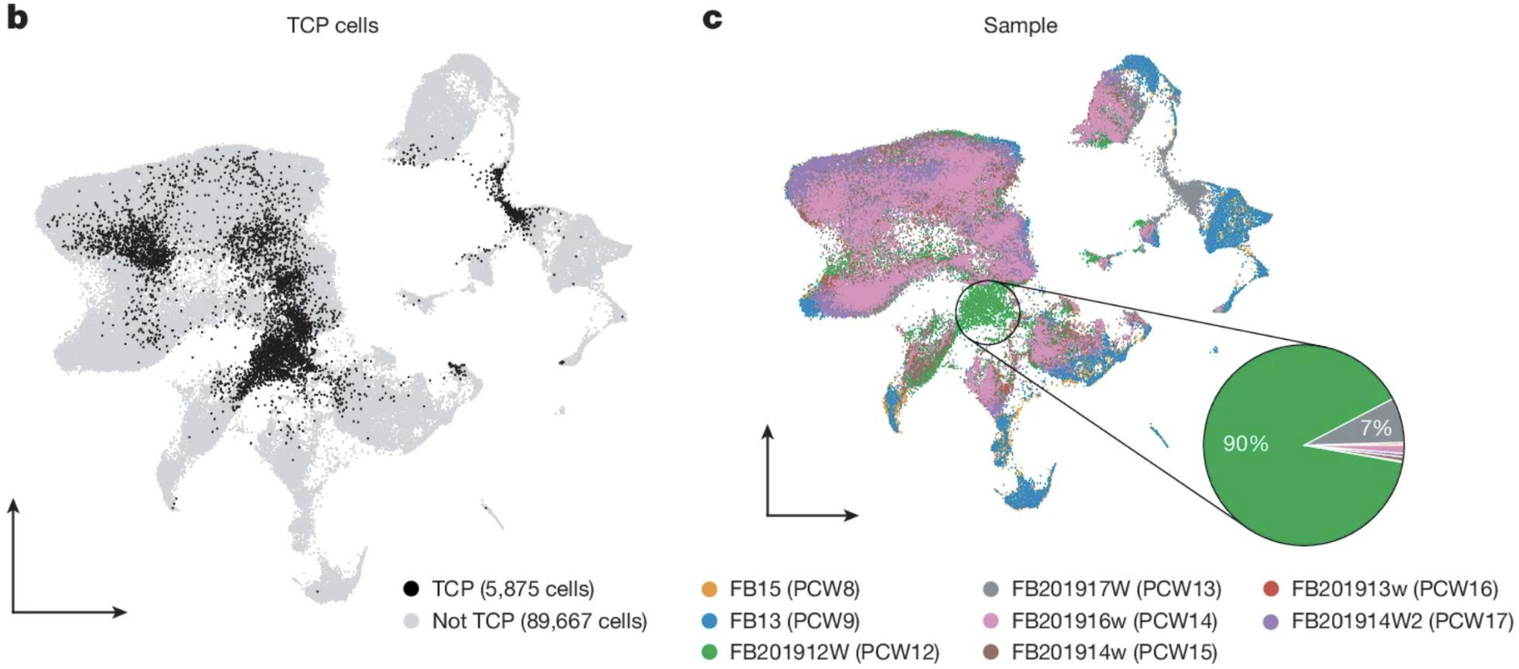
- Look at quality of raw sequencing with fastqs
  - Run FastQC, review sequencing metrics
- Look at quality of alignment
  - Review mapping quality, number of cells, genes, etc.
- See DNB Bioinformatics Core trainings for additional info

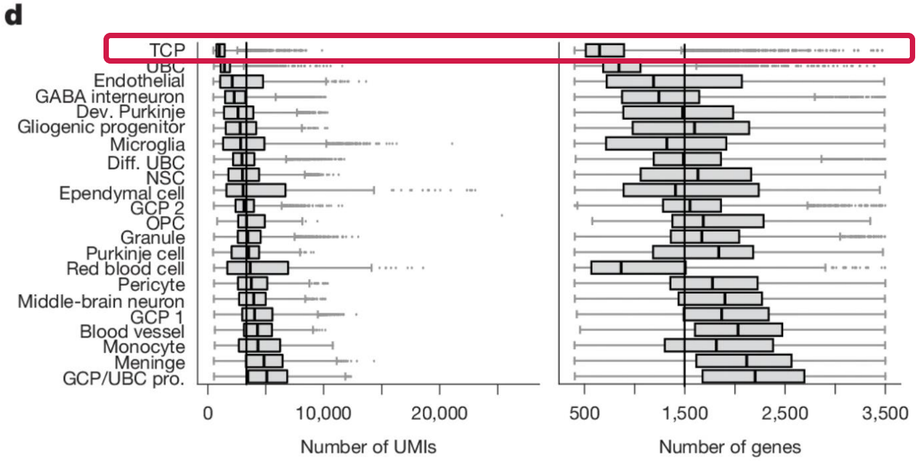
Intro to sc/snRNA-seq workshop

Cell Level QC and Filtering

- First step after alignment, always filter cells
- Low quality cells not informative of true biology
- Skews results, interpretation, can lead to false conclusions
- Cutoffs depend on specific project, data type, sequencing depth
- Choosing appropriate cutoffs is important
    - Too low, keep low quality cells
    - Too high, lose informative cells

Cell Level QC and Filtering



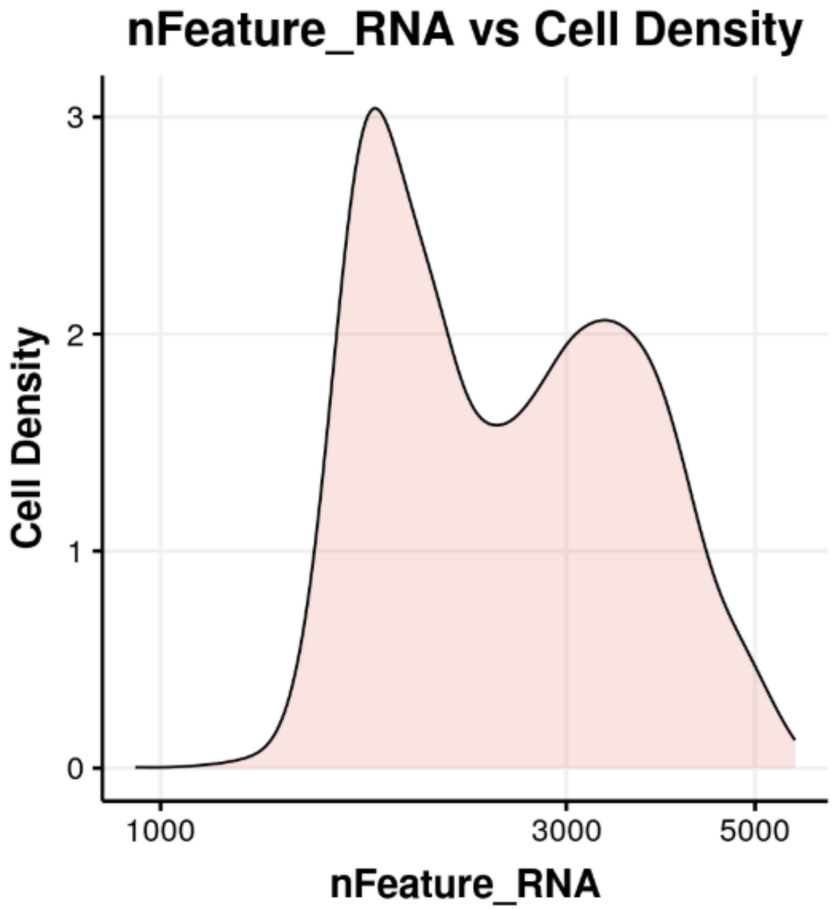


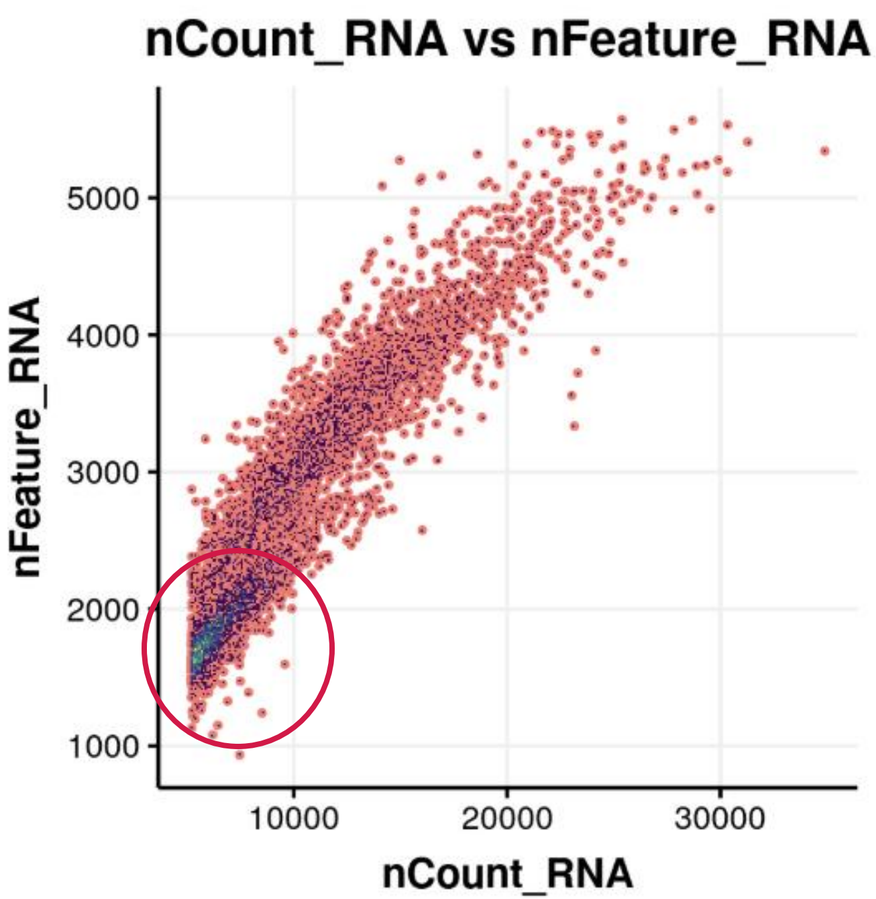
Smith (2025). Nature.

Cell Level QC and Filtering

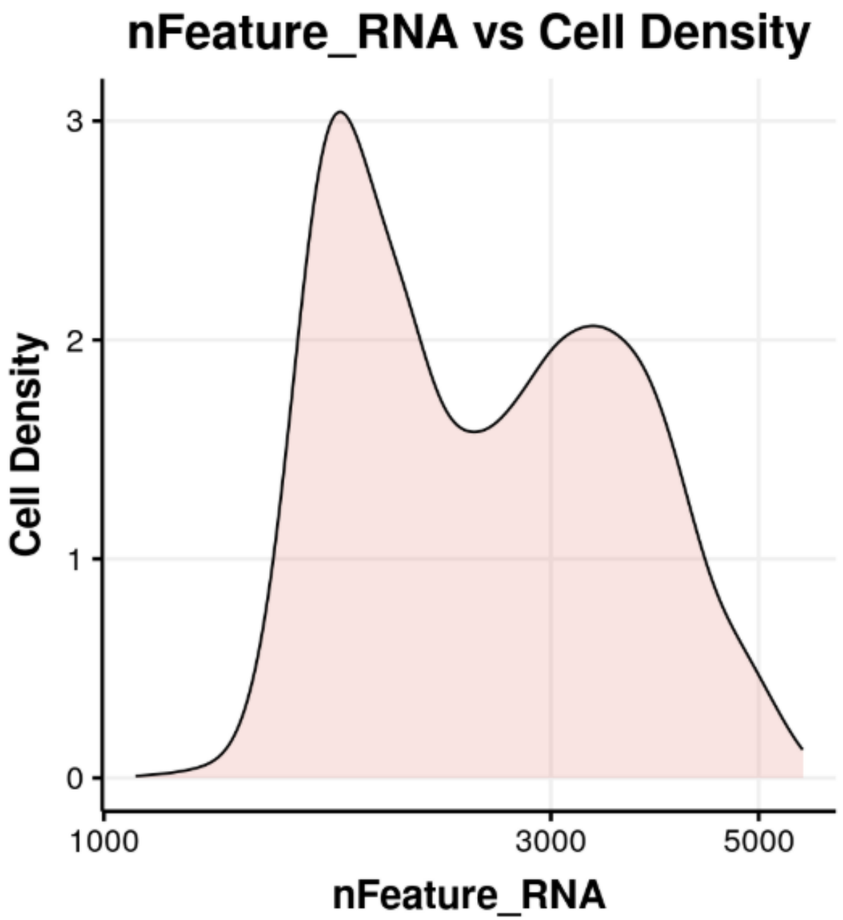
- Minimum: at least 300 genes, at least 500 UMI, less than 5% (scRNA-seq) or 1% (snRNA-seq) mitochondrial expression
- Consider applying outlier filters, e.g. median absolute deviation (MAD)
- Apply filtering to samples individually
- Test multiple parameter sets
- Search literature for cutoffs used for similar cell types, models
- Look at distributions, other plots of metrics
    - Bimodal distributions, indications of lower diversity, cells grouping on filtering metrics in UMAP

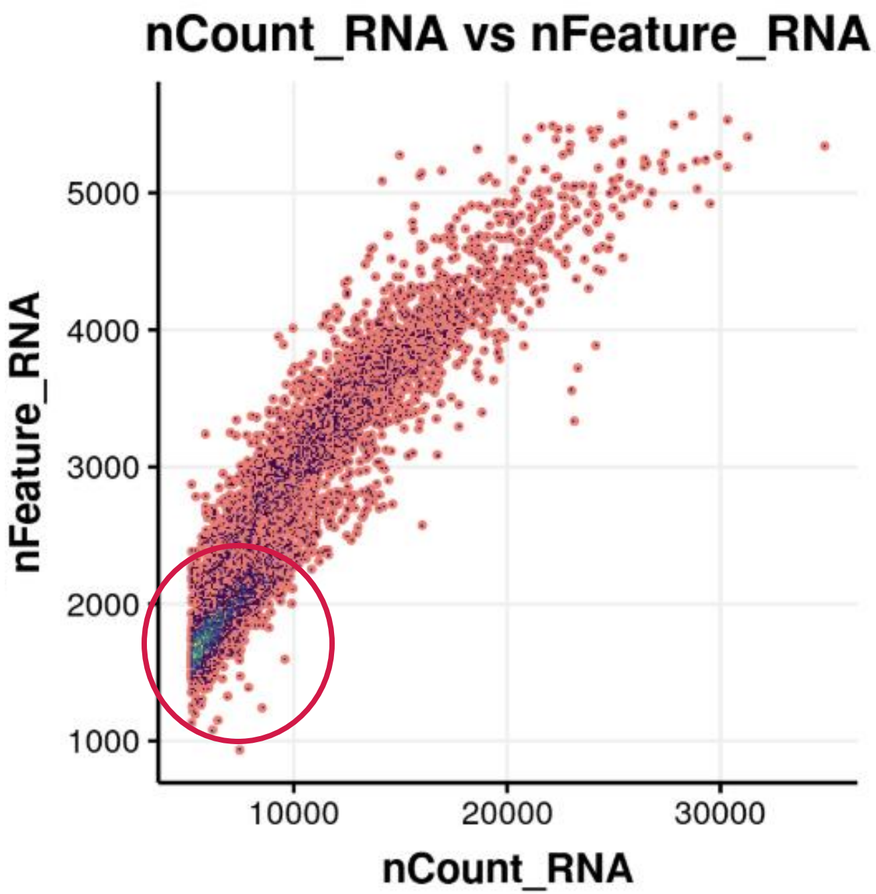
Example: 300 gene + 500 UMI, n cells = 4,837



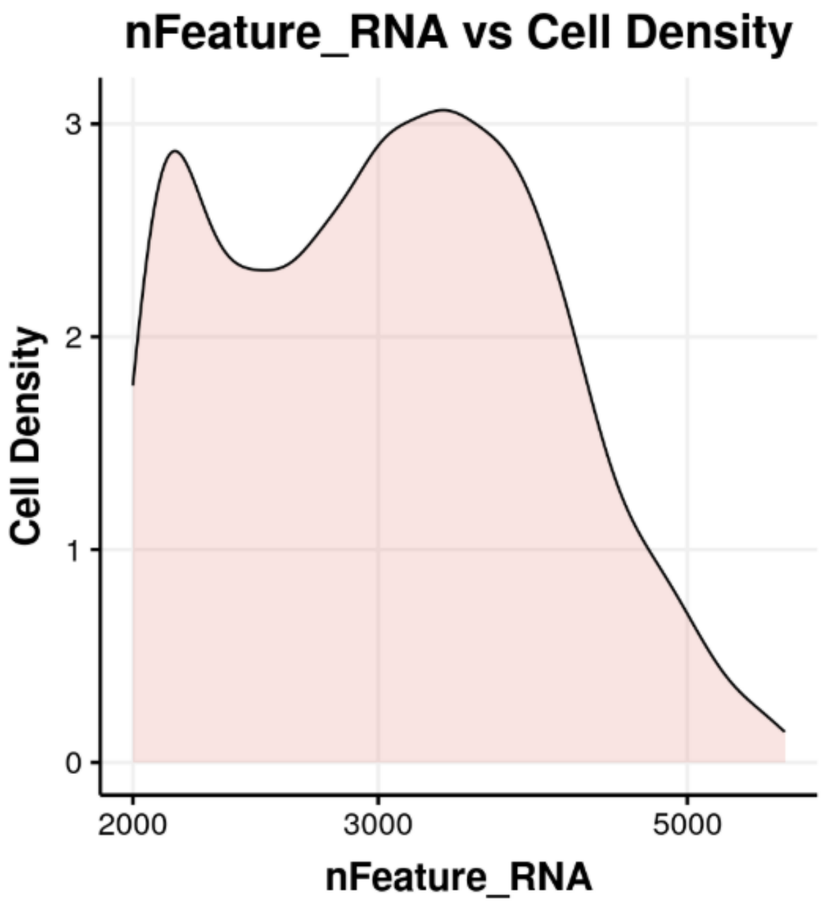


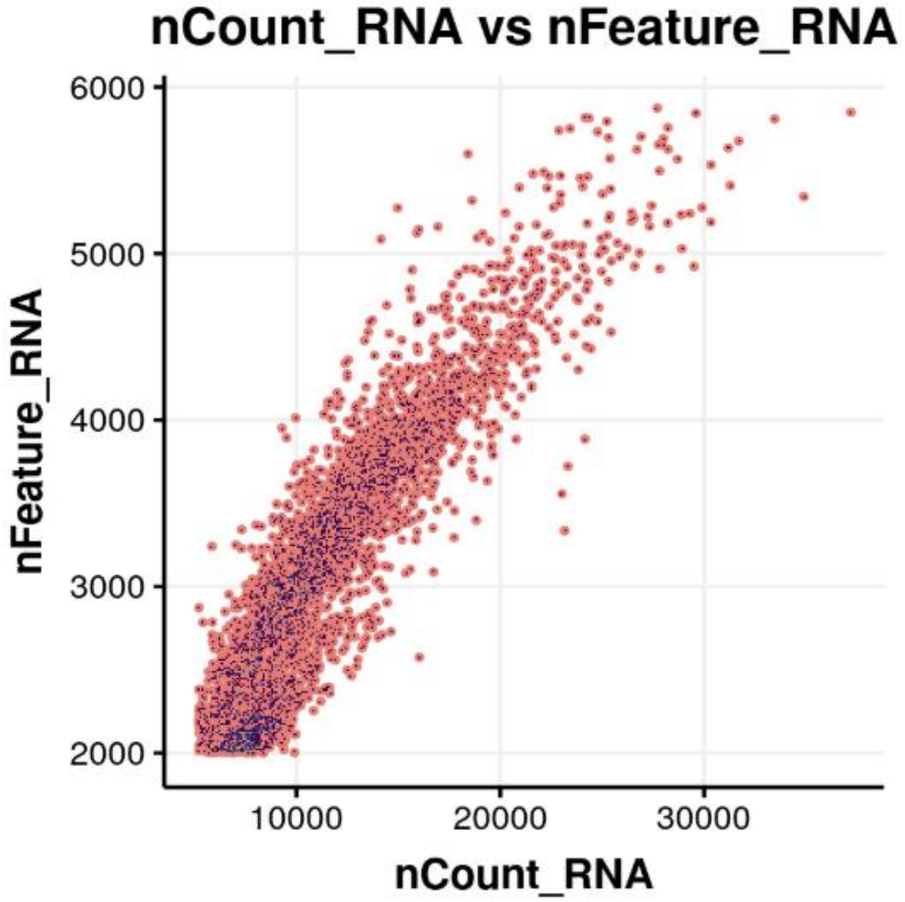
1k gene + 2k UMI, n cells = 4,836





2k gene + 3k UMI, n cells = 3,279

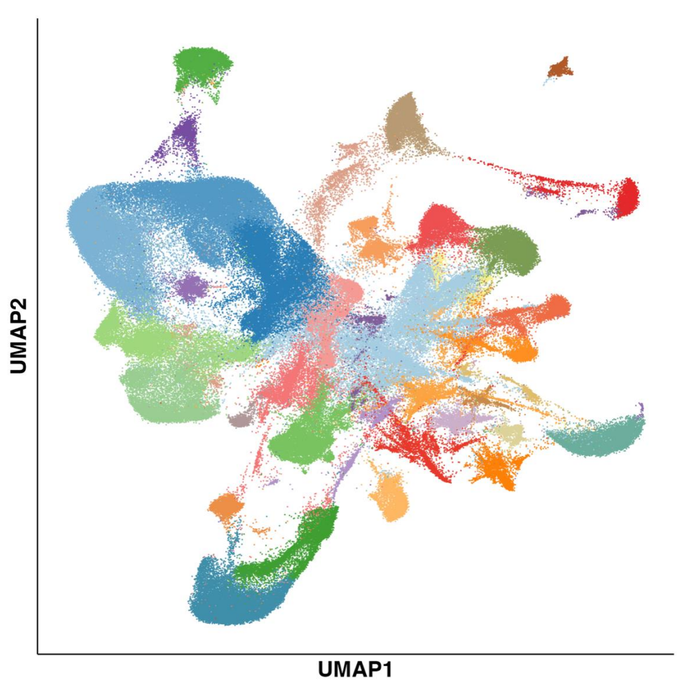


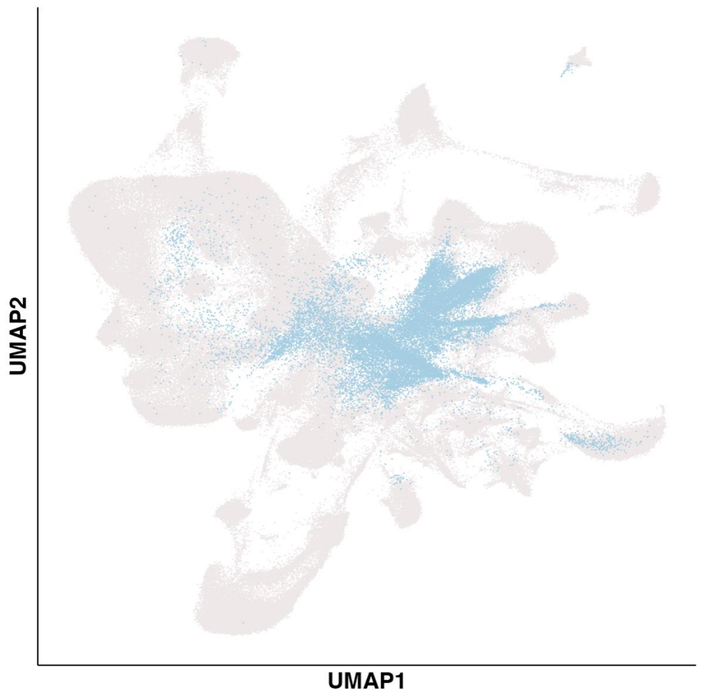


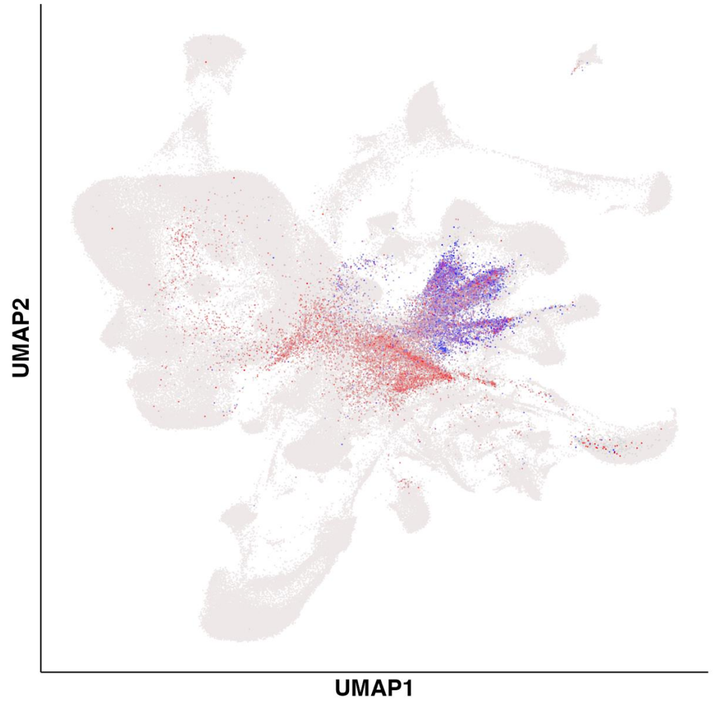
Cell Level QC and Filtering

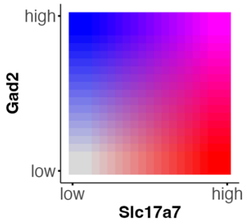
- May need to correct for ambient RNA contamination
- May need to filter out doublets
- Do not recommend correcting, filtering initially
- Tools can struggle with differentiating cell types, overlapping expression
- Patterns in UMAPs, expression could indicate contamination, doublets
    - Spiderwebbing (ambient RNA contamination)
    - Lack of distinct clusters
    - Overlap in mutually exclusive marker expression

Cell Level QC and Filtering

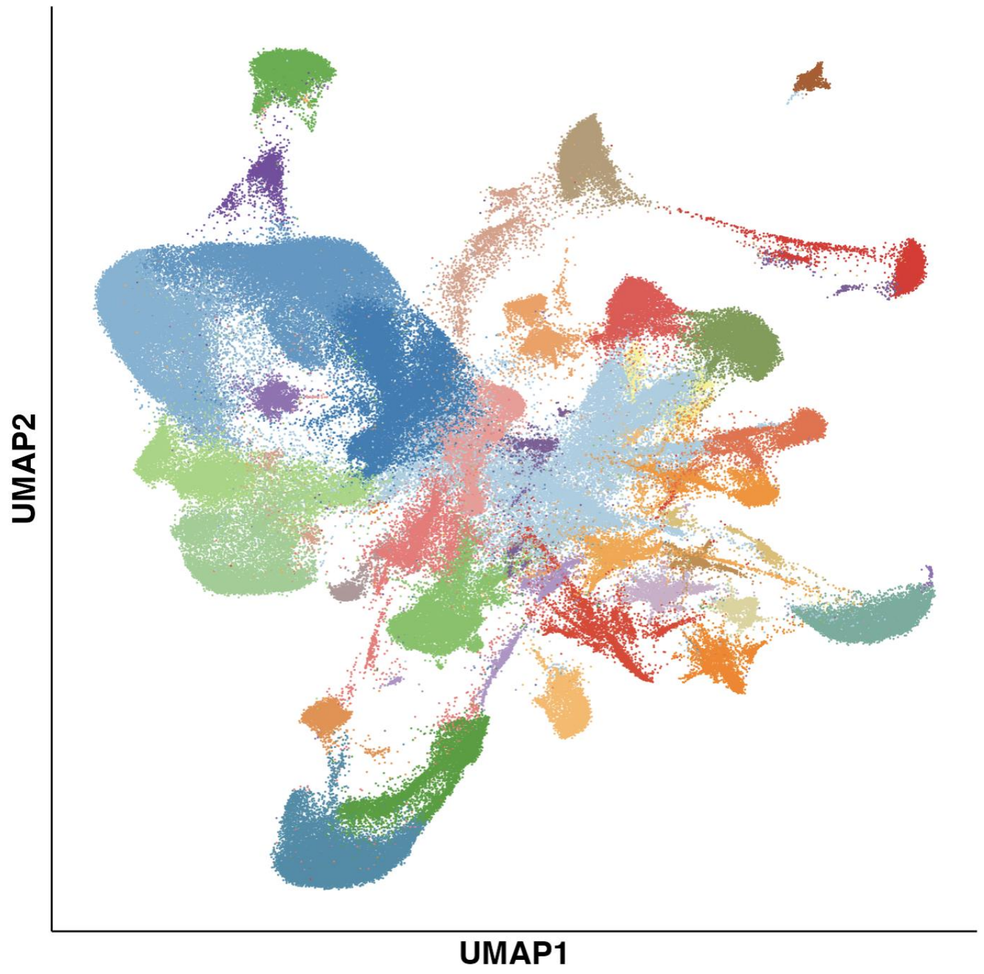


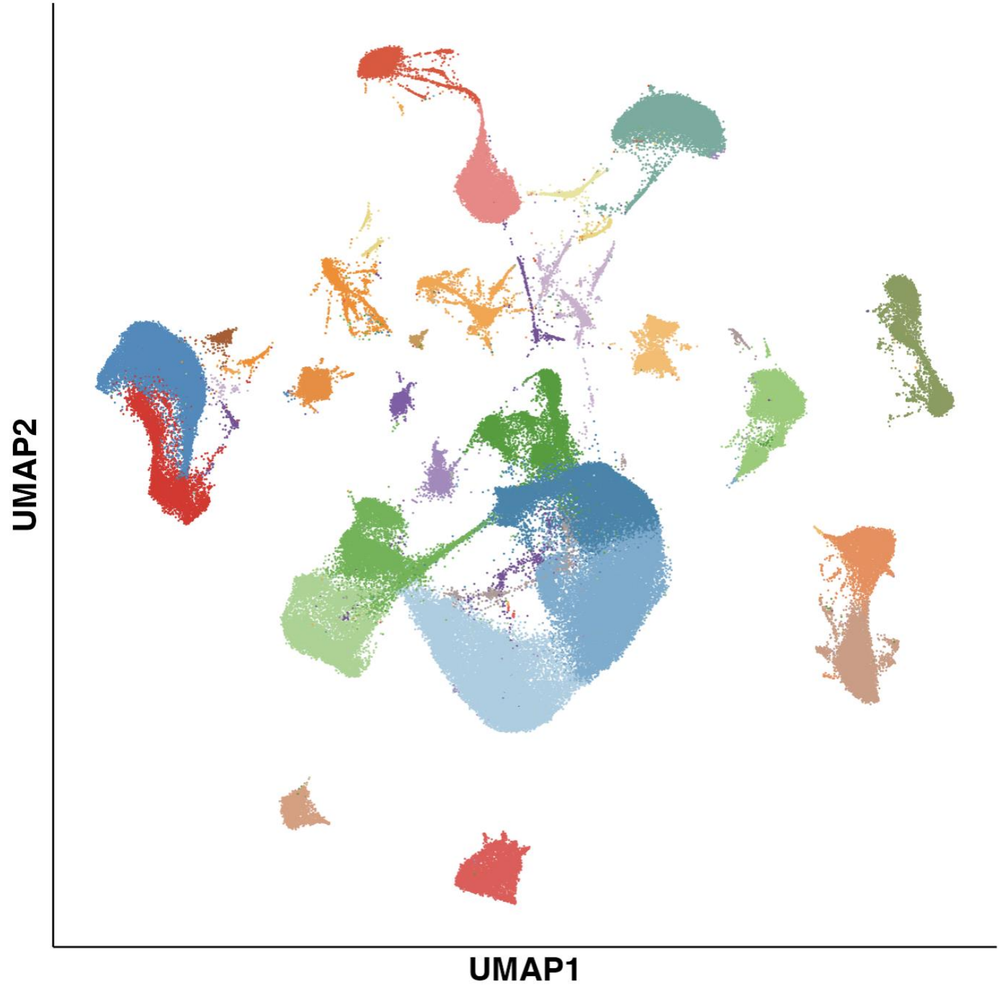






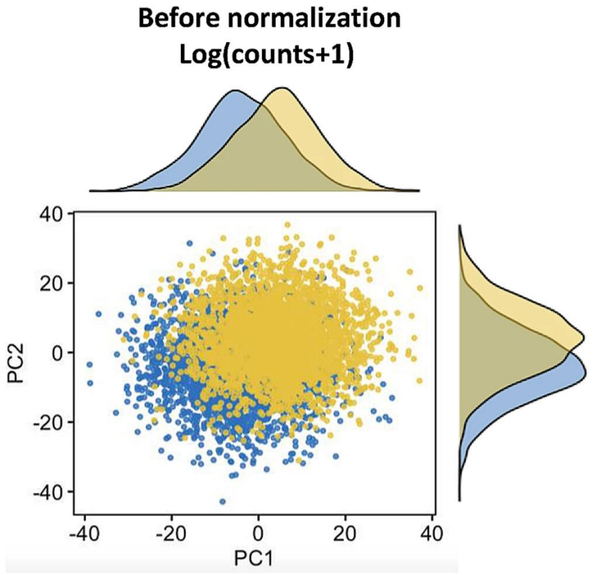
Cell Level QC and Filtering

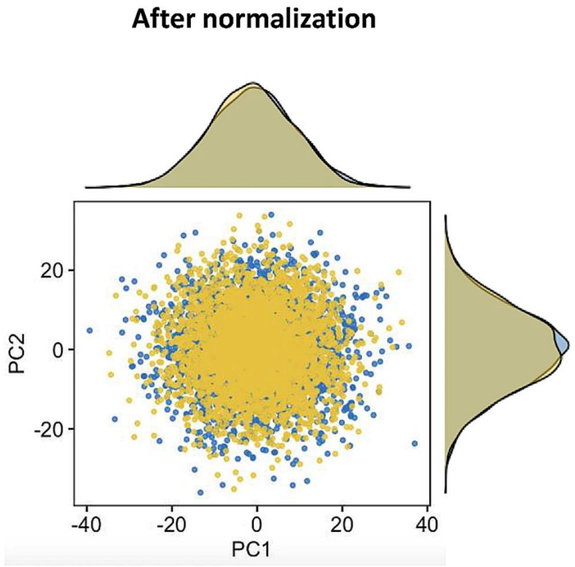




Normalization

- Important to make gene counts comparable
  - Relevant within and across samples
  - Reduces sources of technical noise
- Seurat has several methods:
  - NormalizeData()
    - LogNormalize (default)
    - Relative Counts
    - Center Log Ratio Transformation
  - SCTransform()
- Must apply the same method across samples





Cuevas-Diaz Duran (2024). BMC Genomics.

Highly Variable Gene (HVG) Selection

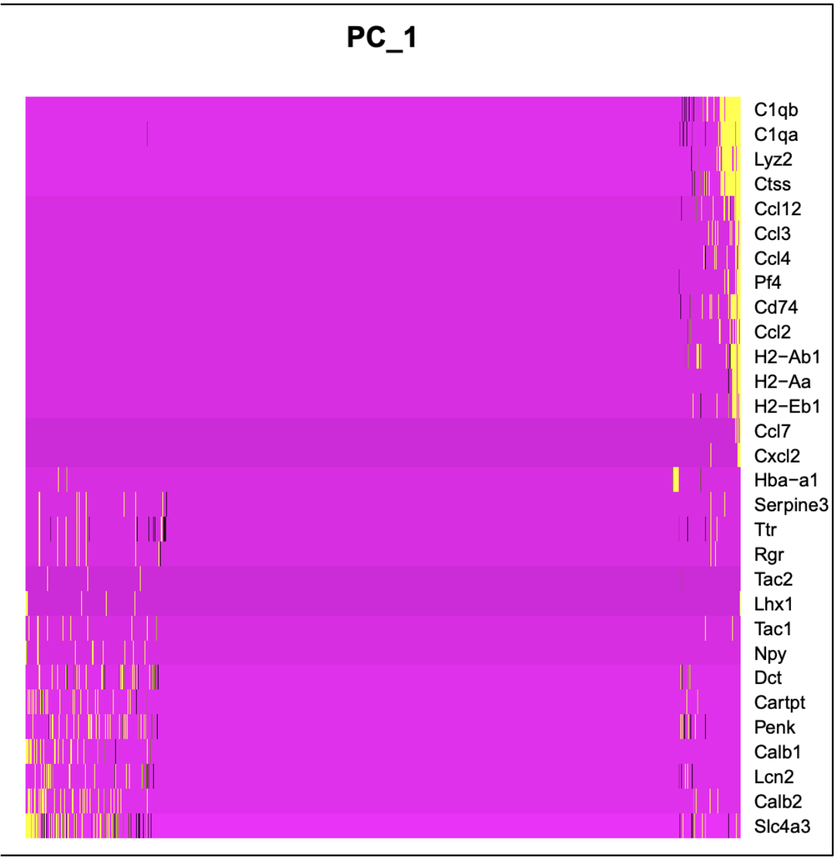
- Identify genes highly expressed in some cells, lowly expressed in others
- HVGs impact downstream results
  - Used for principal component analysis (PCA)
  - PCs used for clustering, non-linear dimensionality reduction, etc.
- Focus on HVGs shown to highlight biological signals
- Feature selection method is important
  - Variance alone may not reflect biology
  - Consider mean-variance relationship

Highly Variable Gene (HVG) Selection

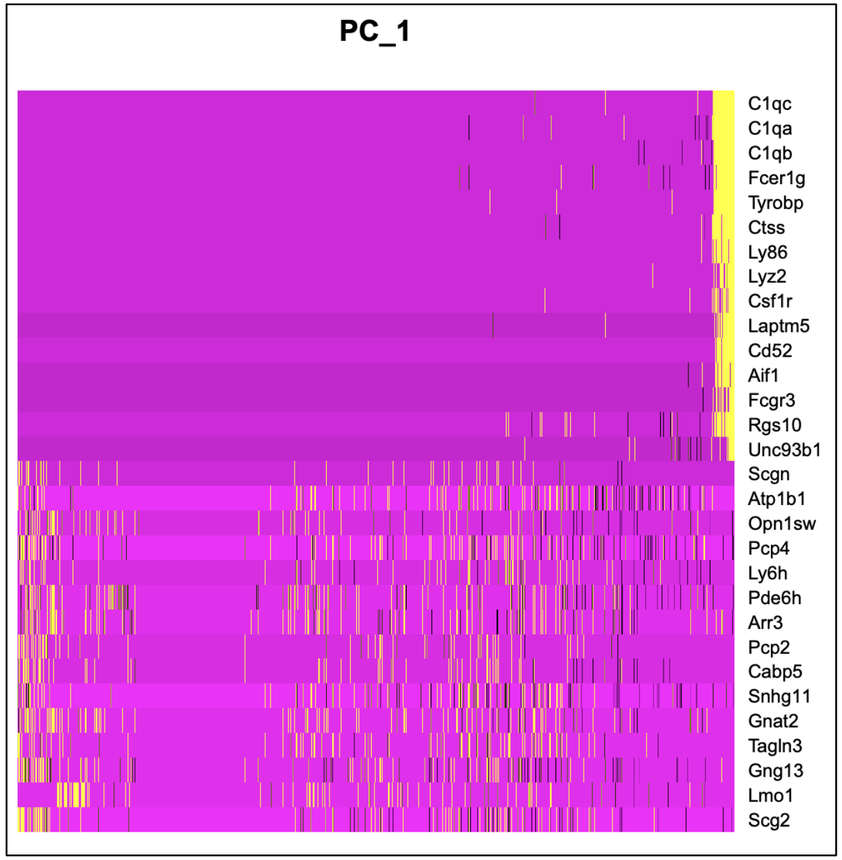
- Seurat default is 2000
- DNB Bioinformatics core default typically 3000
- Won’t necessarily change this parameter
- Be aware number of HVGs impacts downstream results
- Always rerun after merging multiple samples

Highly Variable Gene (HVG) Selection

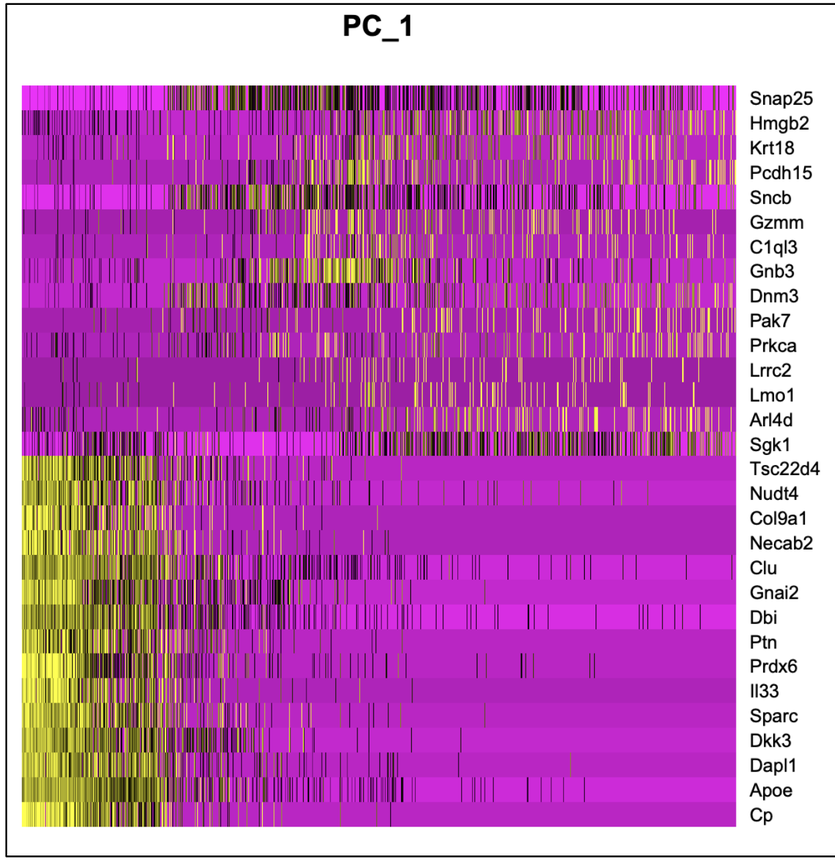
of HVGs = 30



of HVGs = 300

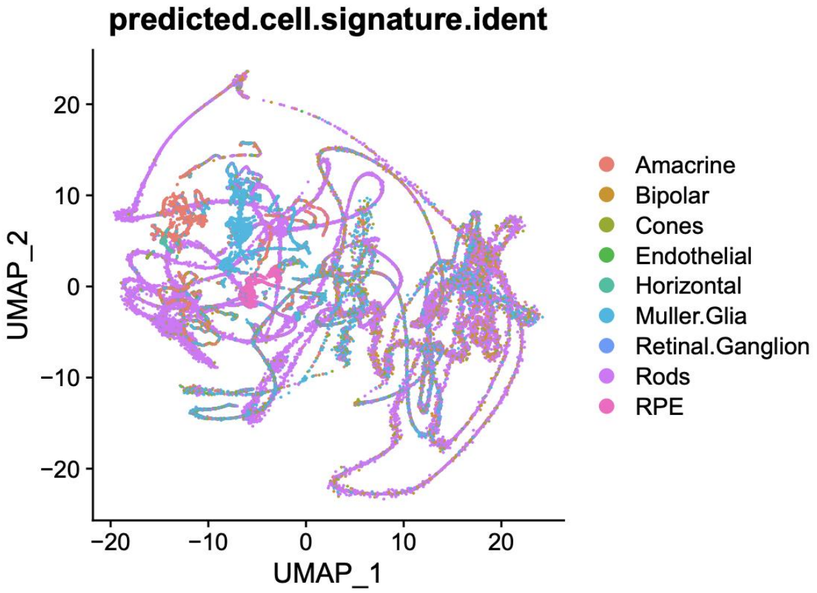


of HVGs = 3000

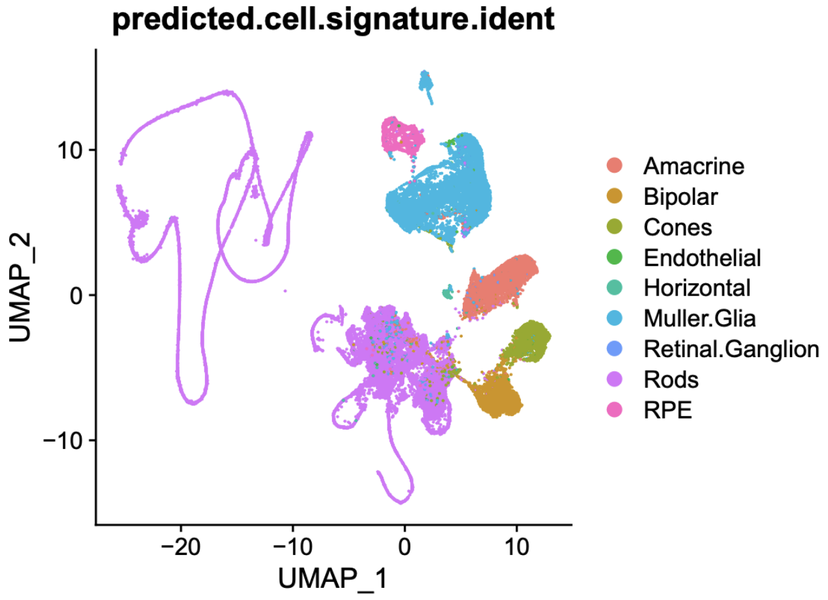


Highly Variable Gene (HVG) Selection

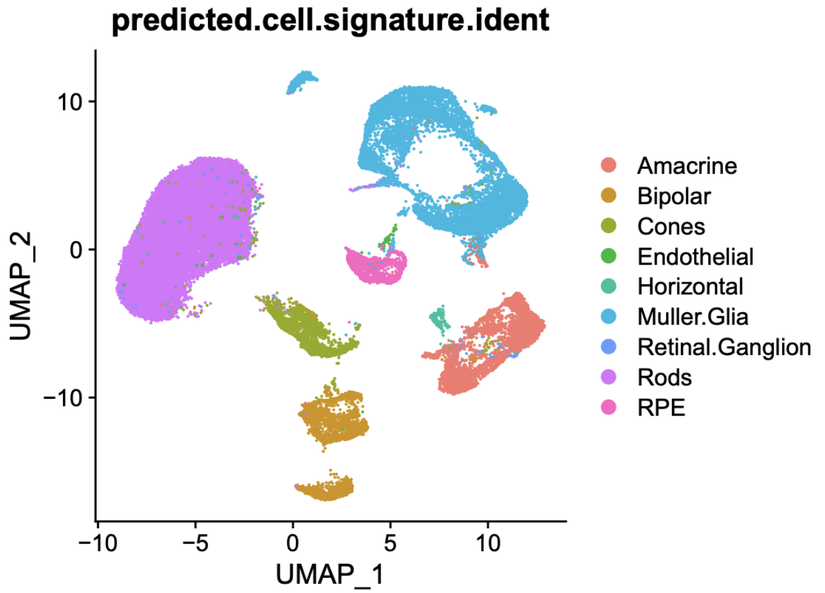
of HVGs = 30



of HVGs = 300



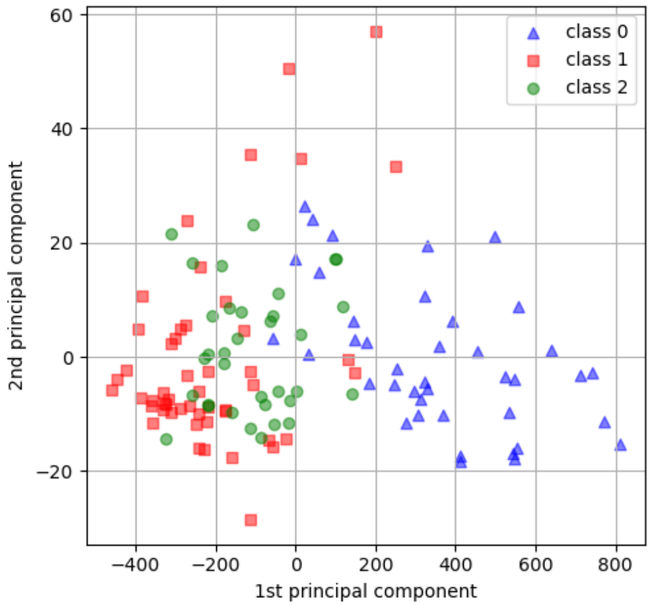
of HVGs = 3000



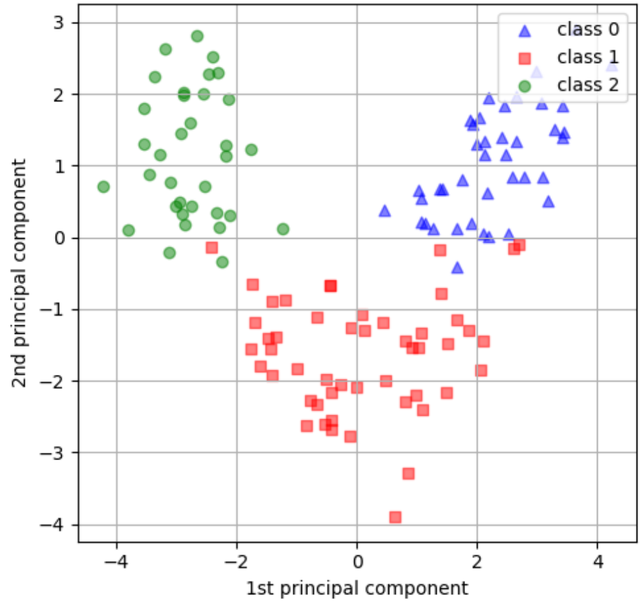
Scaling

- Scaling performed per gene
- Makes gene expression more comparable across cells, samples
- Reduces dominance of highly expressed genes
- Important for principal component analysis, downstream steps

PCA on unscaled data



PCA on scaled data



[Importance of Feature Scaling.]

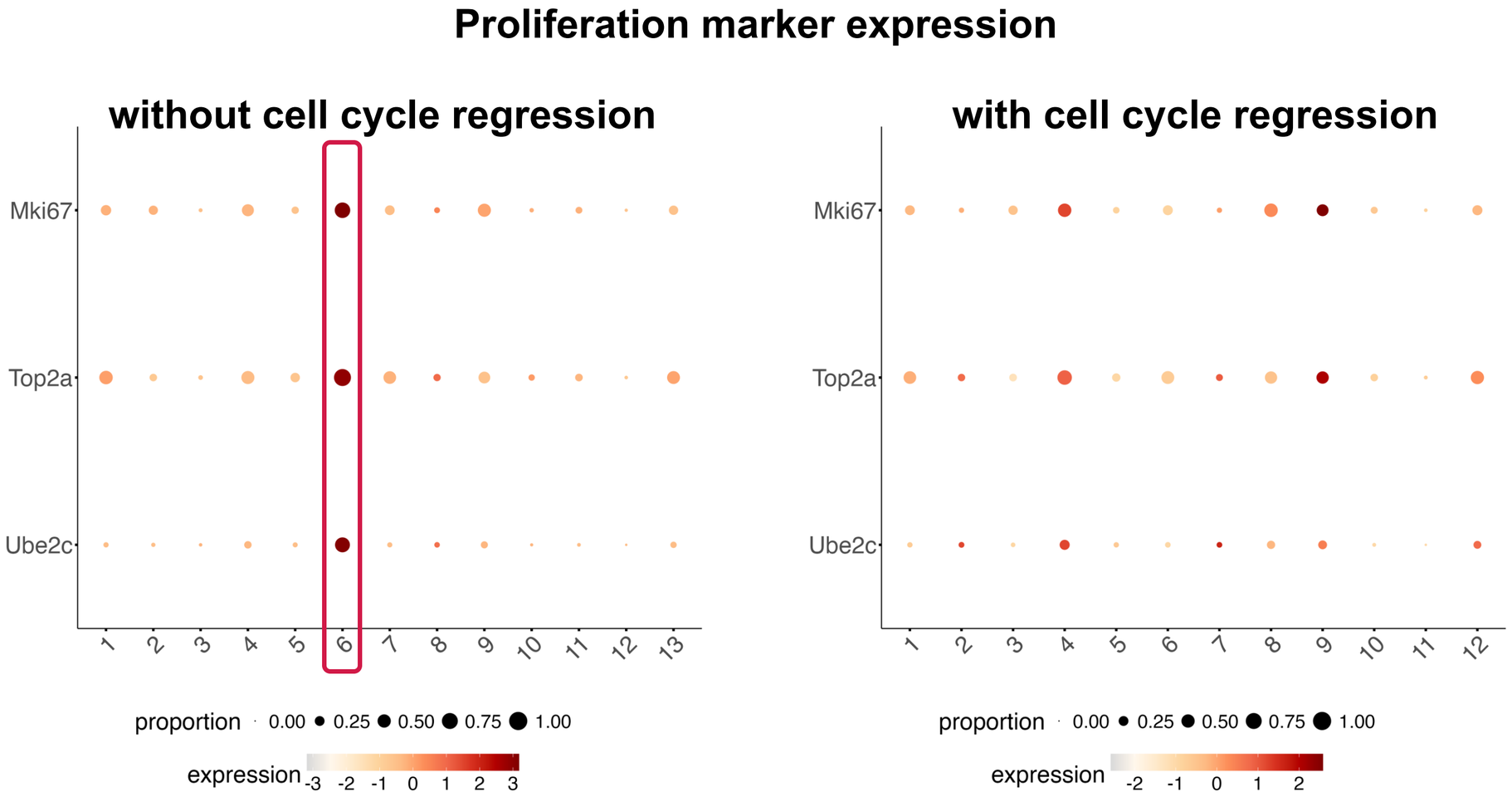
Scaling

- Center and scale expression of each gene
  - Mean = 0
  - Standard deviation = 1
  - z-score transformation
- Scale before running PCA
- Always scaling rerun after merging samples
- Not rescaling can contribute to batch effects, technical artefacts

Scaling and Regression

- Can do regression during scaling step
- Regress out technical or biological variables
  - Cell cycle genes, mitochondrial genes, ribosomal genes, UMI count, batch
- Do analysis without regression first, add regression later if needed
- Seurat has three models for regression
  - Linear (default), Poisson (GLM), Negative Binomial (GLM)
- Regression can have unintended effects
- If regressing multiple variables, regress in same step

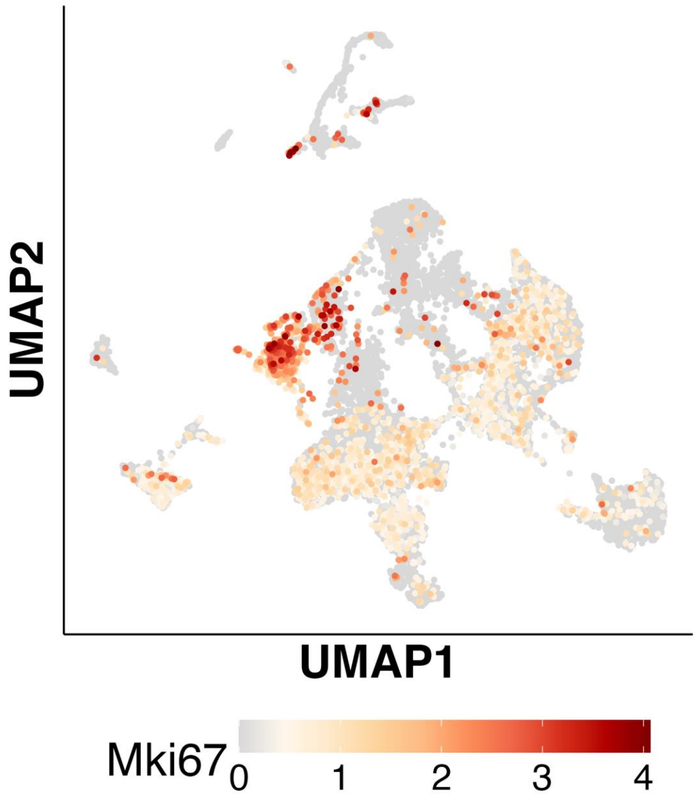
Scaling and Regression



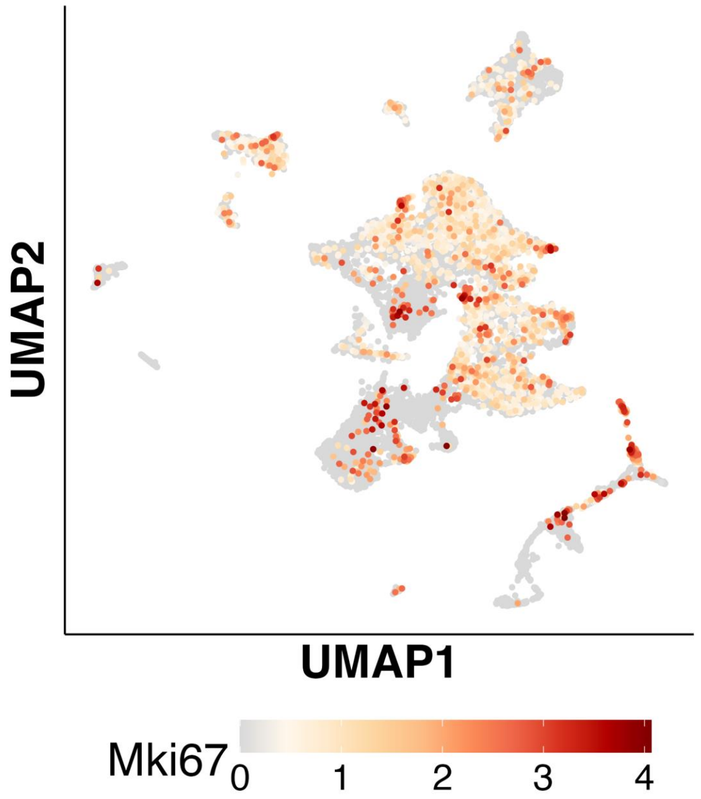
Scaling and Regression

Proliferation marker expression

without cell cycle regression



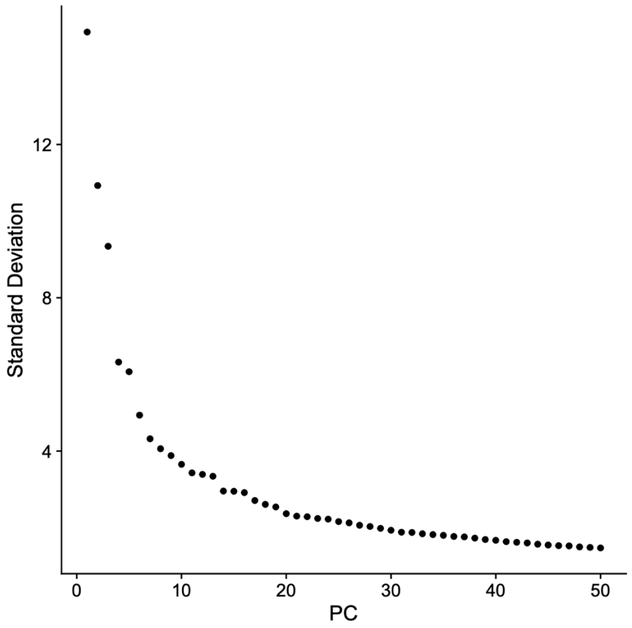
with cell cycle regression



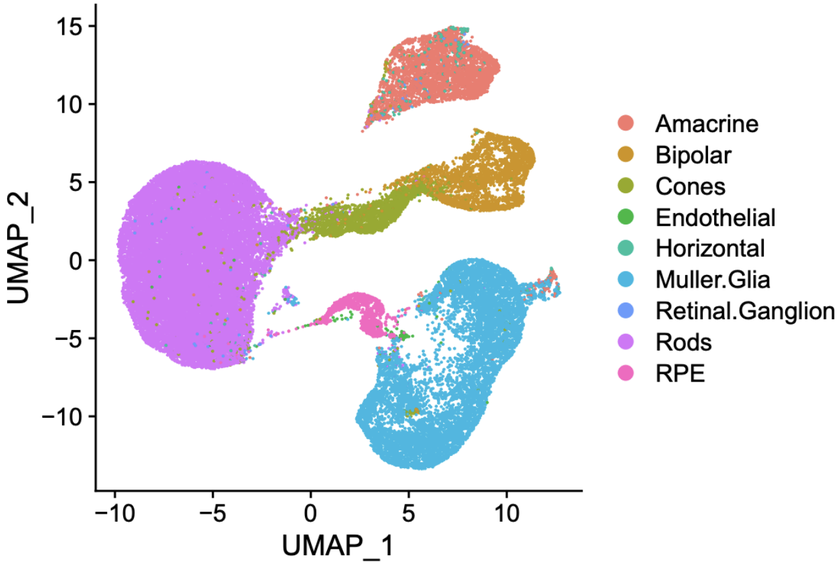
PCA (linear dimensionality reduction)

- Principal component analysis (PCA) is common linear dimensionality reduction
  - Identify gene sets with correlated expression patterns
  - Capture heterogeneity across cells with fewer dimensions
- Choose an appropriate number of PCs
  - Too few, won’t distinguish biology
  - Too many, include technical noise, have diminishing returns
- Sufficient range typically 10 to 50 PCs
- Use elbow plots for choosing reasonable number

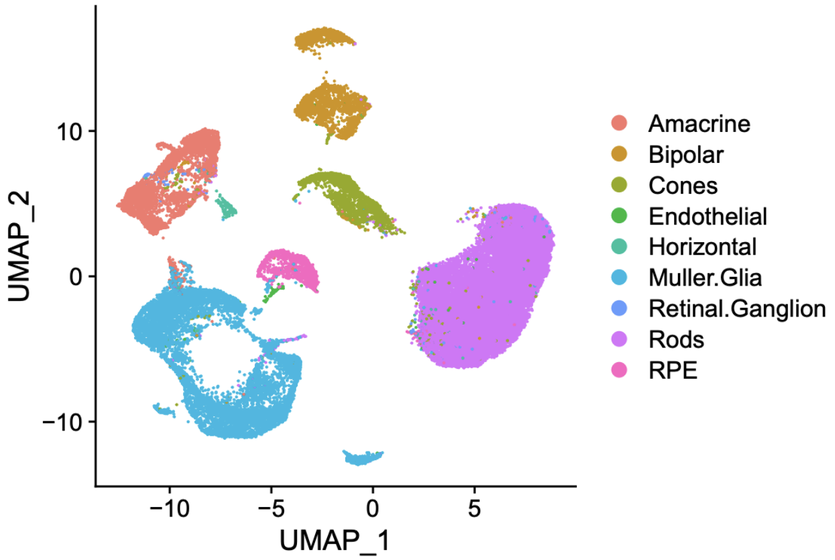
Elbow plot



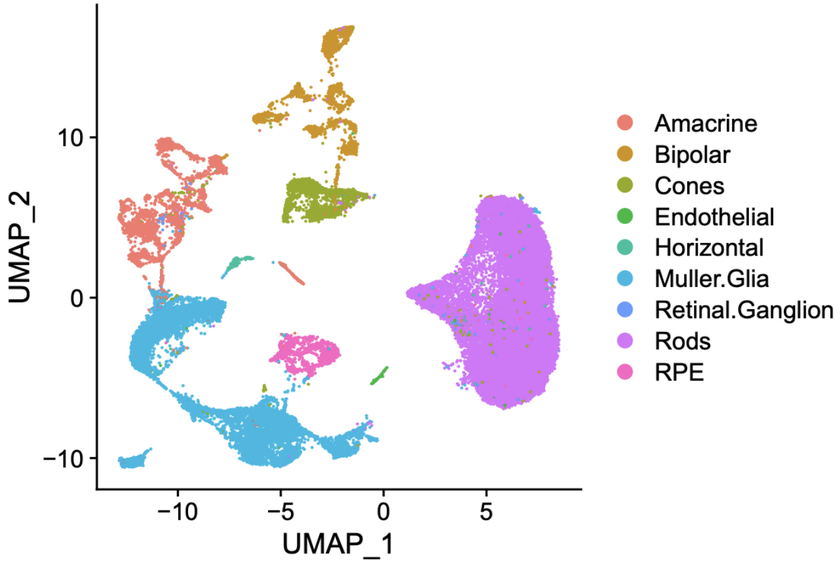
of PCs = 5



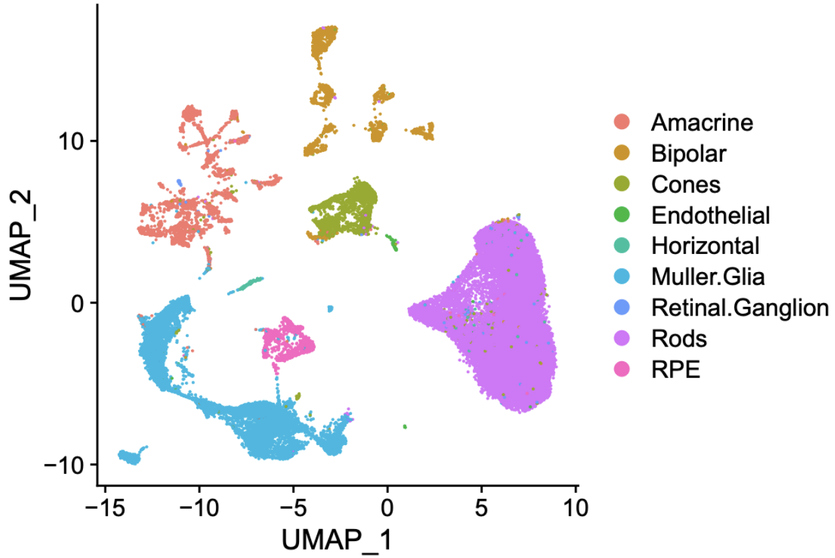
of PCs = 10



of PCs = 30

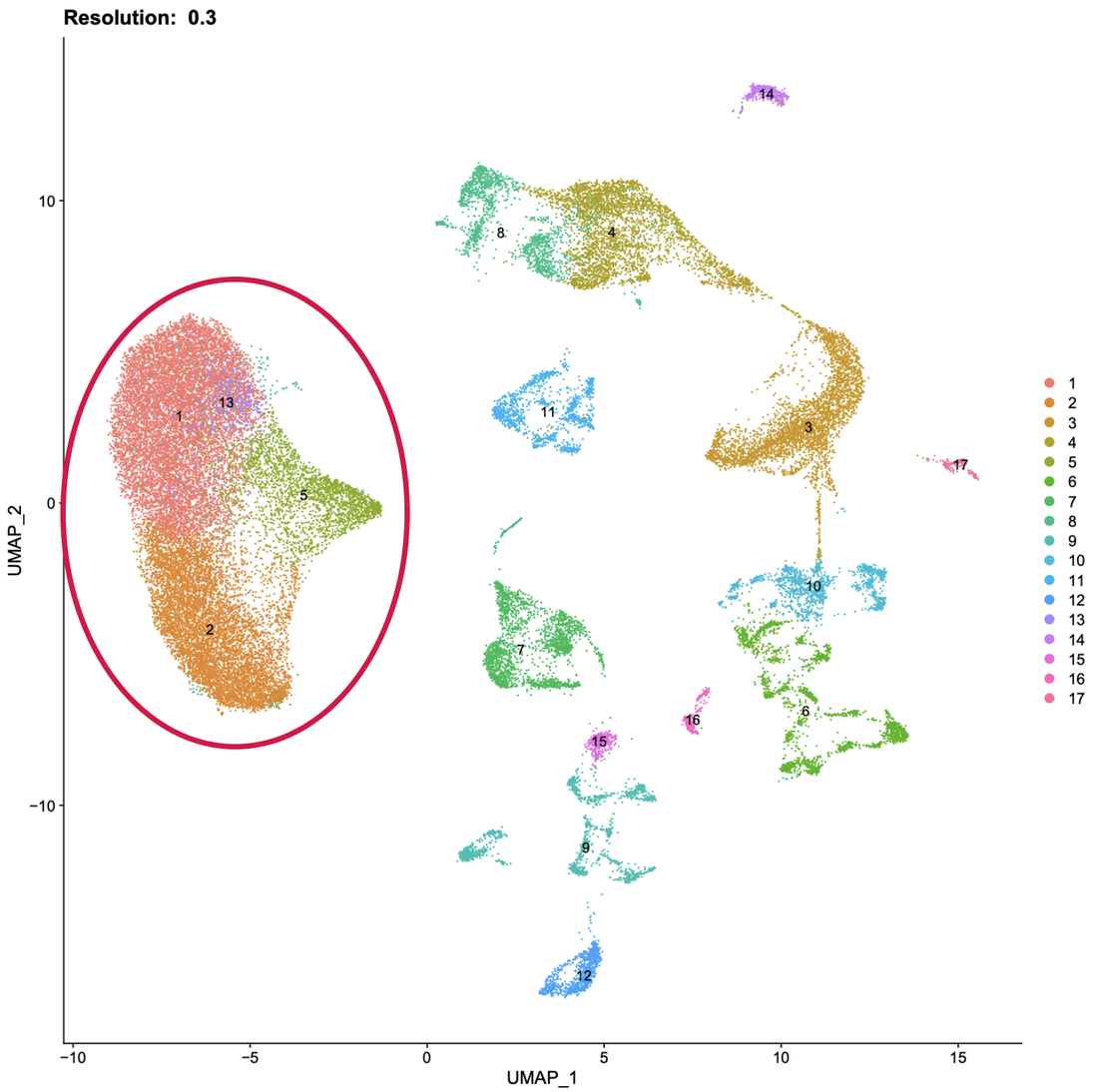


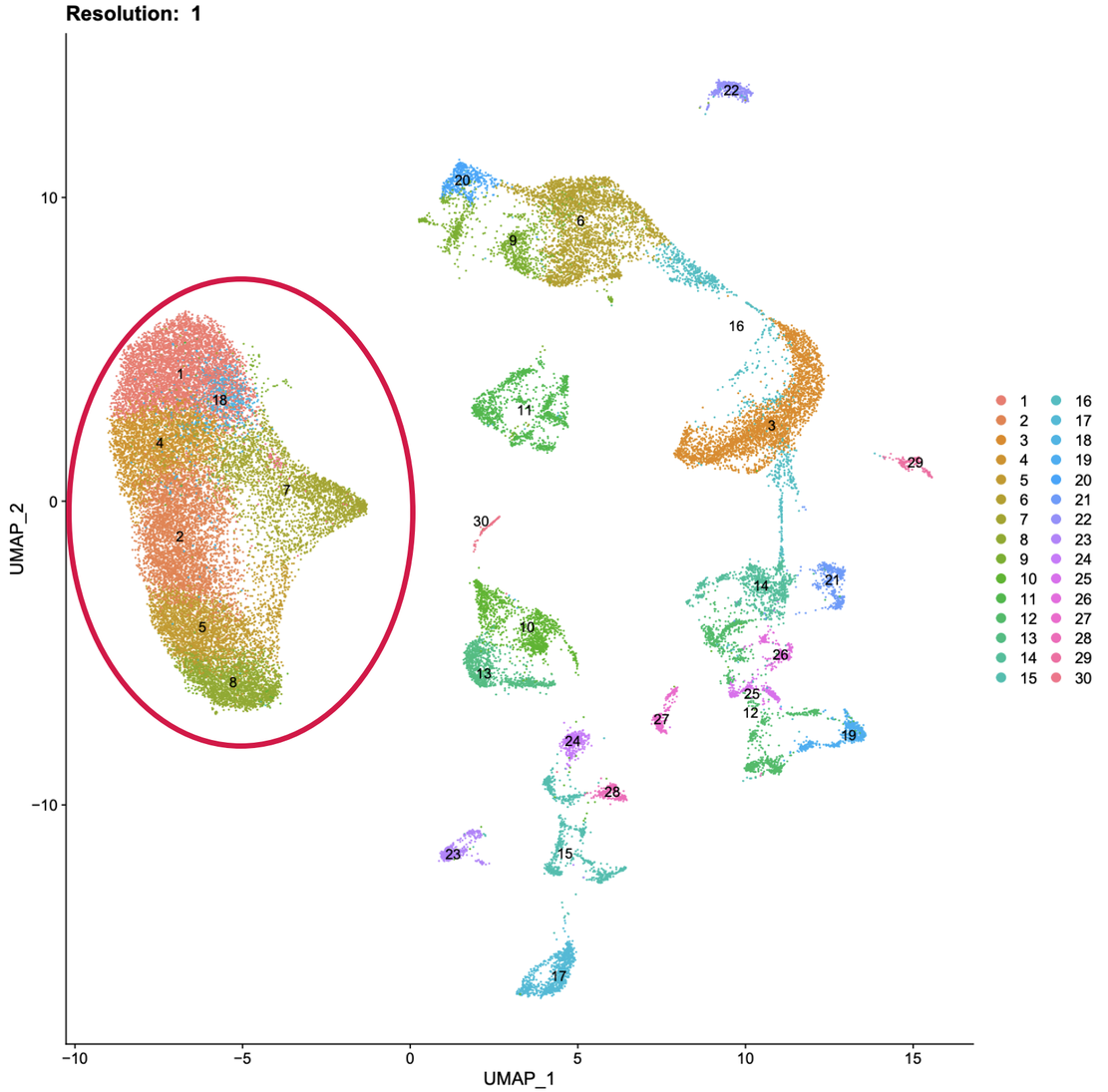
of PCs = 50



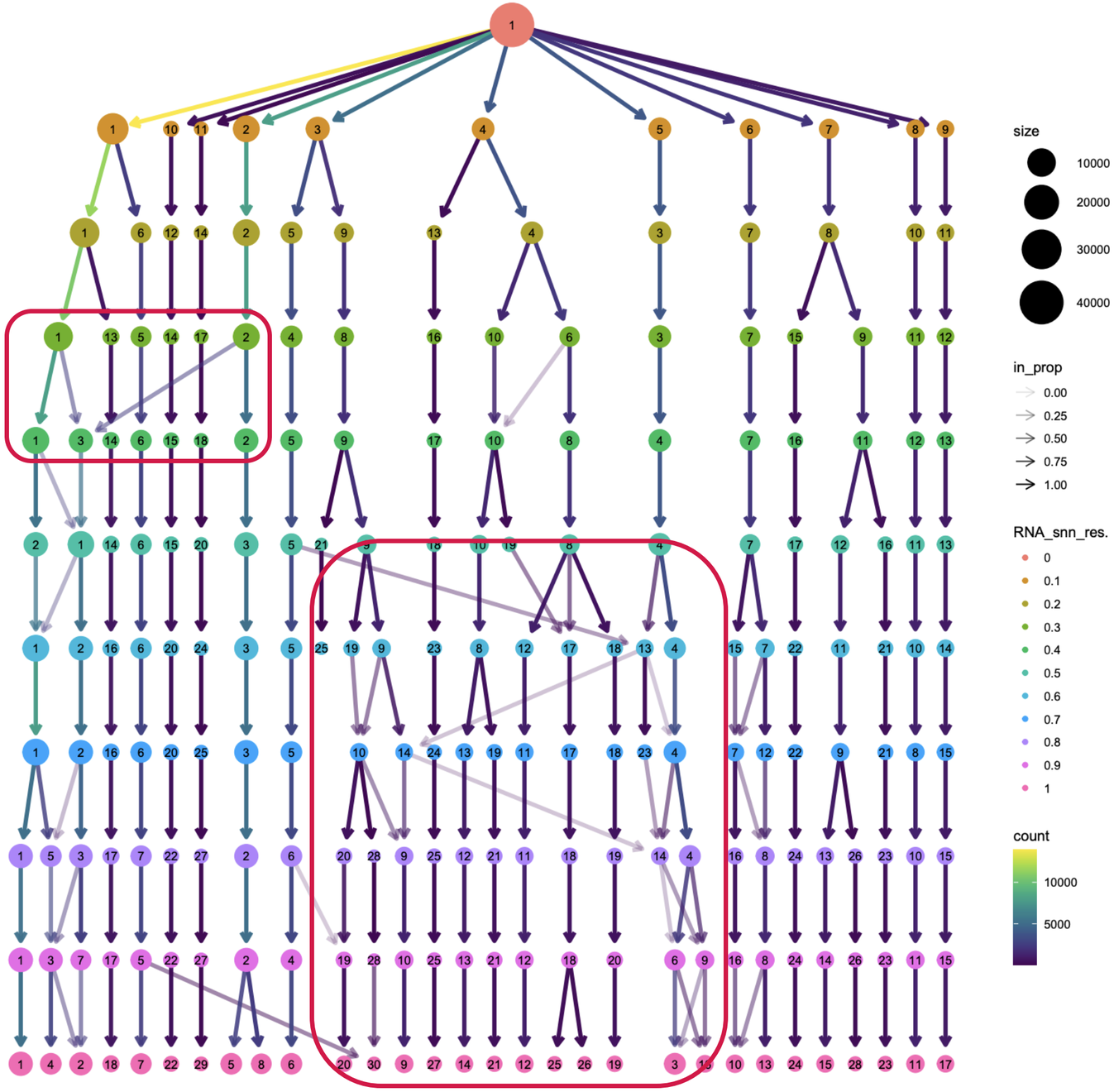
Clustering

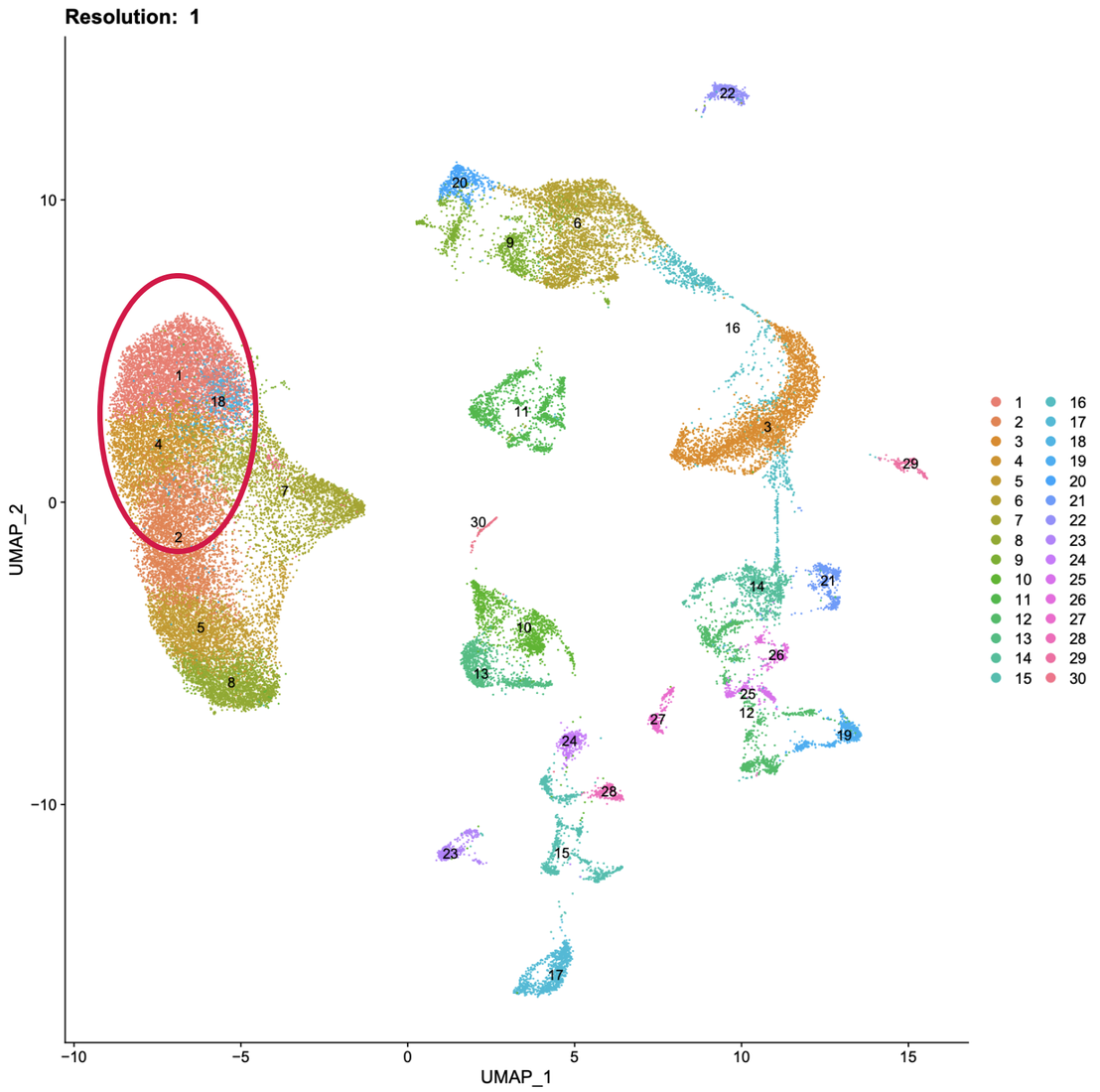
- Generating clusters is common next step after PCA
- Two steps in Seurat
  - FindNeighbors() – uses PCs, determines overlap in close cells
  - FindClusters() – uses FindNeighbors() similarity scores, groups cells optimally
- Important parameter for FindClusters() is resolution
- Resolution determines how broad or fine clustering is
  - Too low, group heterogenous cells together
  - Too high, split homogenous groups arbitrarily

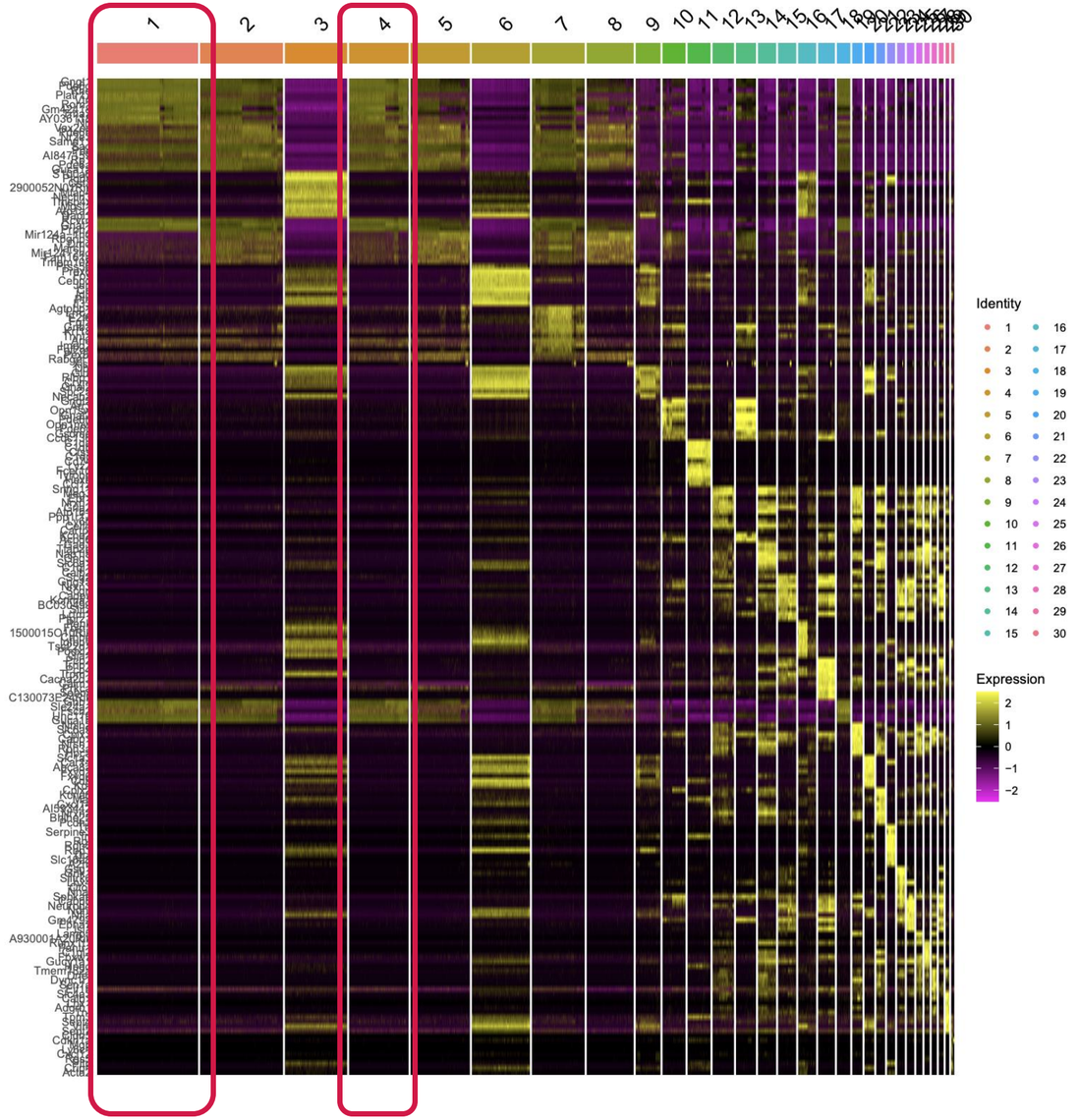

Resolution: 0.3
UMAP_2
UMAP_1
10
0
−10
−10
−5
0
5
10
15
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17


Resolution: 1
UMAP_2
UMAP_1
10
0
−10
−10
−5
0
5
10
15
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30

- Each circle is a cluster.
  - Size represents relative number of cells
- Each level corresponds to resolution.
  - Resolutions 0 to 1, increments of 0.1
- Arrows represent how cells split, move across clusters from one resolution to the next
- Multiple incoming arrows to cluster, cells jumping, not splitting suggest overclustering




Resolution: 1
UMAP_2
UMAP_1
10
0
−10
−10
−5
0
5
10
15
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30


1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
Identity
Expression
2
1
0
−1
−2

Non-linear dimensionality reduction and visualization

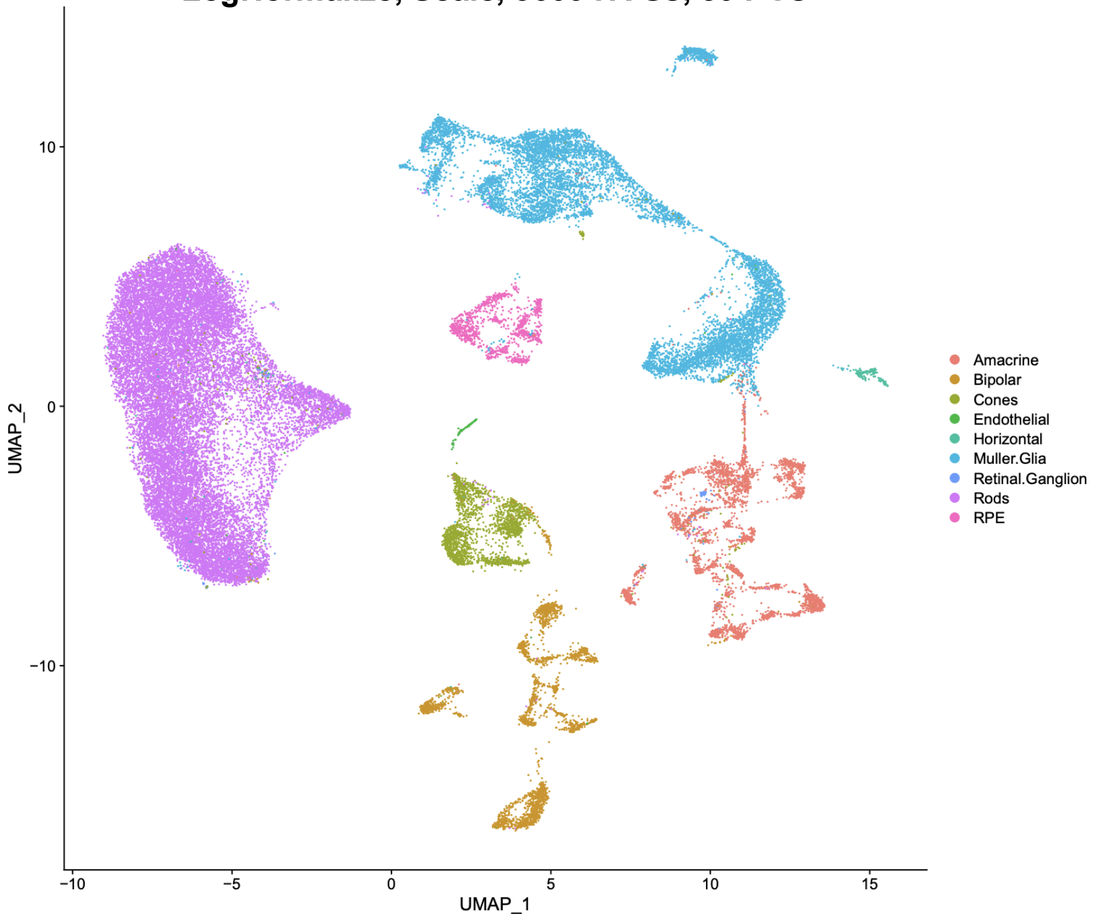
- Non-linear dimensionality reduction for visualization common
  - UMAP
- Display cells in two dimensions, preserve underlying data structure
  - E.g clusters ideally group in UMAP
- Uses PCs, improves speed, loses some structure
- Non-linear dimensionality reduction methods are imperfect.
- 2D cannot capture full picture of true structure

Non-linear dimensionality reduction and visualization

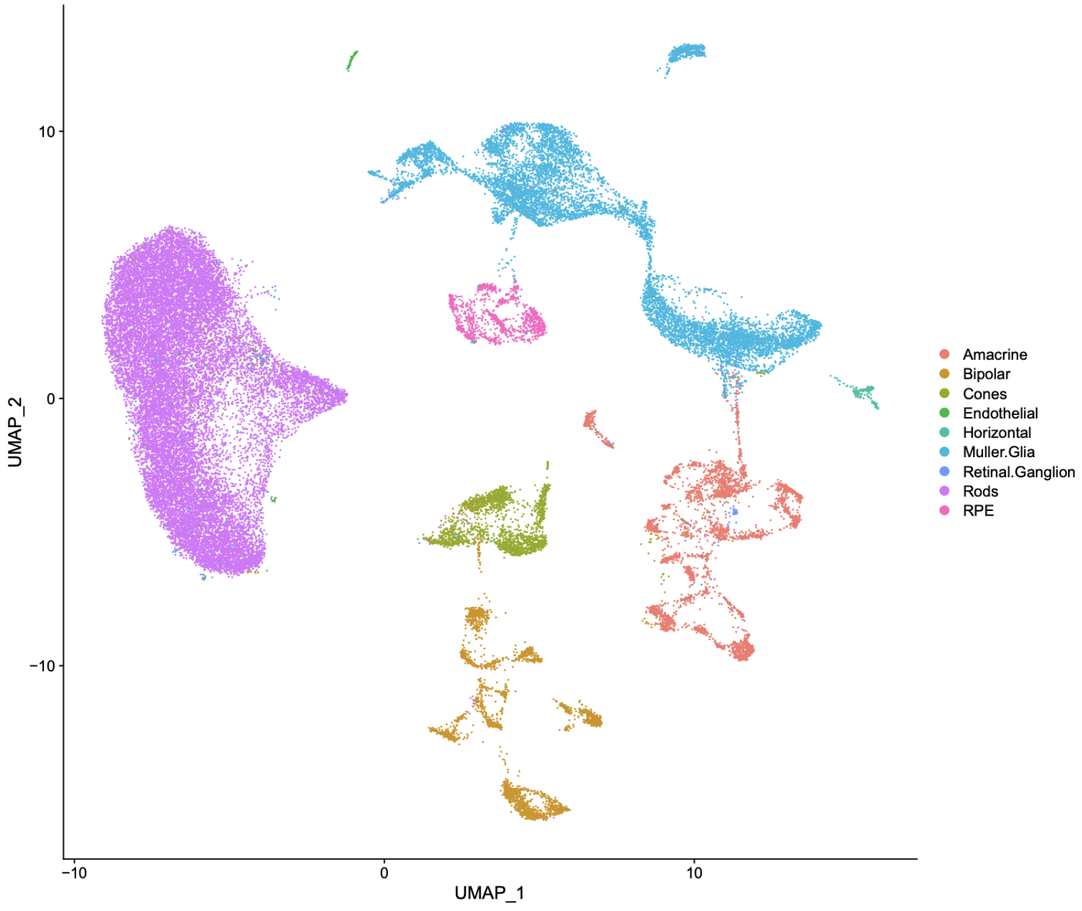
- Visualizations can help you refine data, e.g. patterns may suggest:
  - Ambient RNA contamination, doublets
  - Cells grouping on cell cycle
  - Cells grouping on filtering metrics
- Do not rely on UMAPs for biological interpretation.
- Shape can easily change based on numerous factors
  - Seed, normalization, scaling, variable features, principal components
- Structure may match biology, but verify through additional methods
- Don’t force structure to fit biology

Non-linear dimensionality reduction and visualization

**“Standard” processing –
LogNormalize, Scale, 3000 HVGs, 30 PCs**

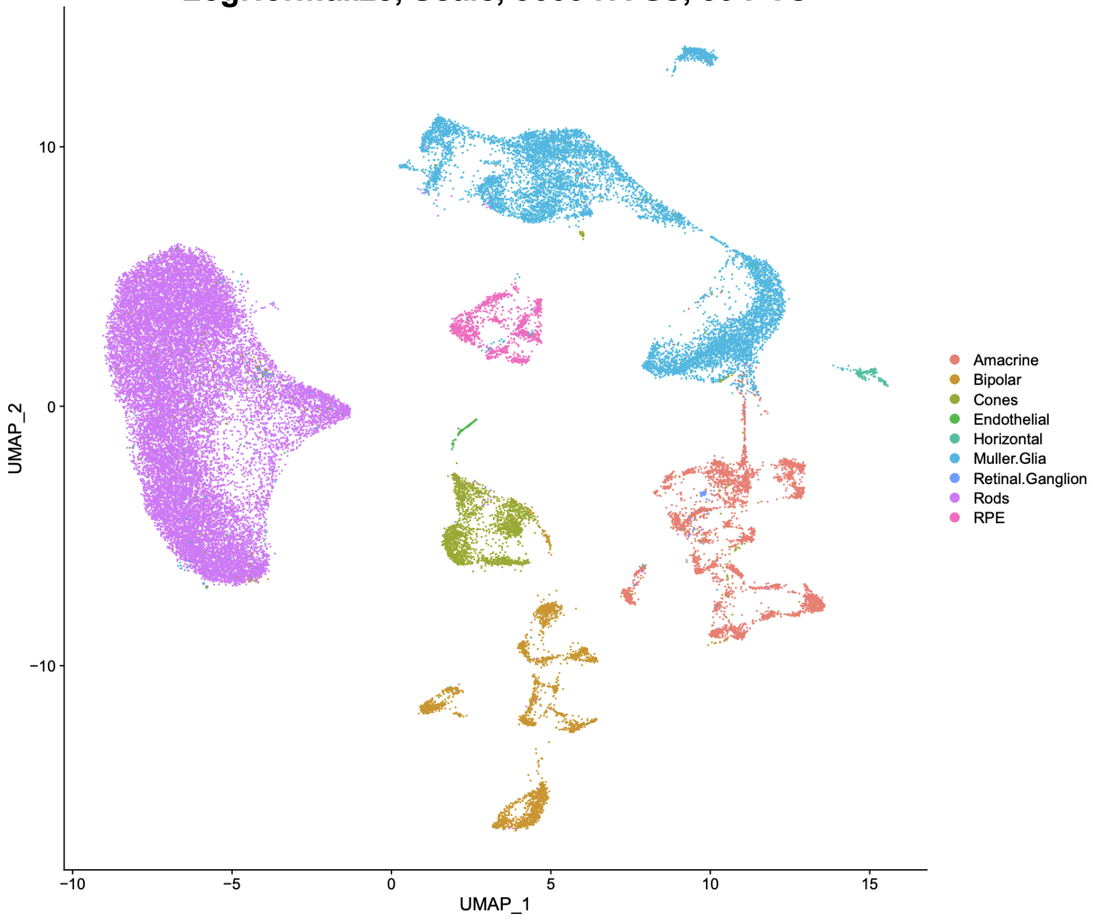


**“Standard” processing –
different seed value (22465)**

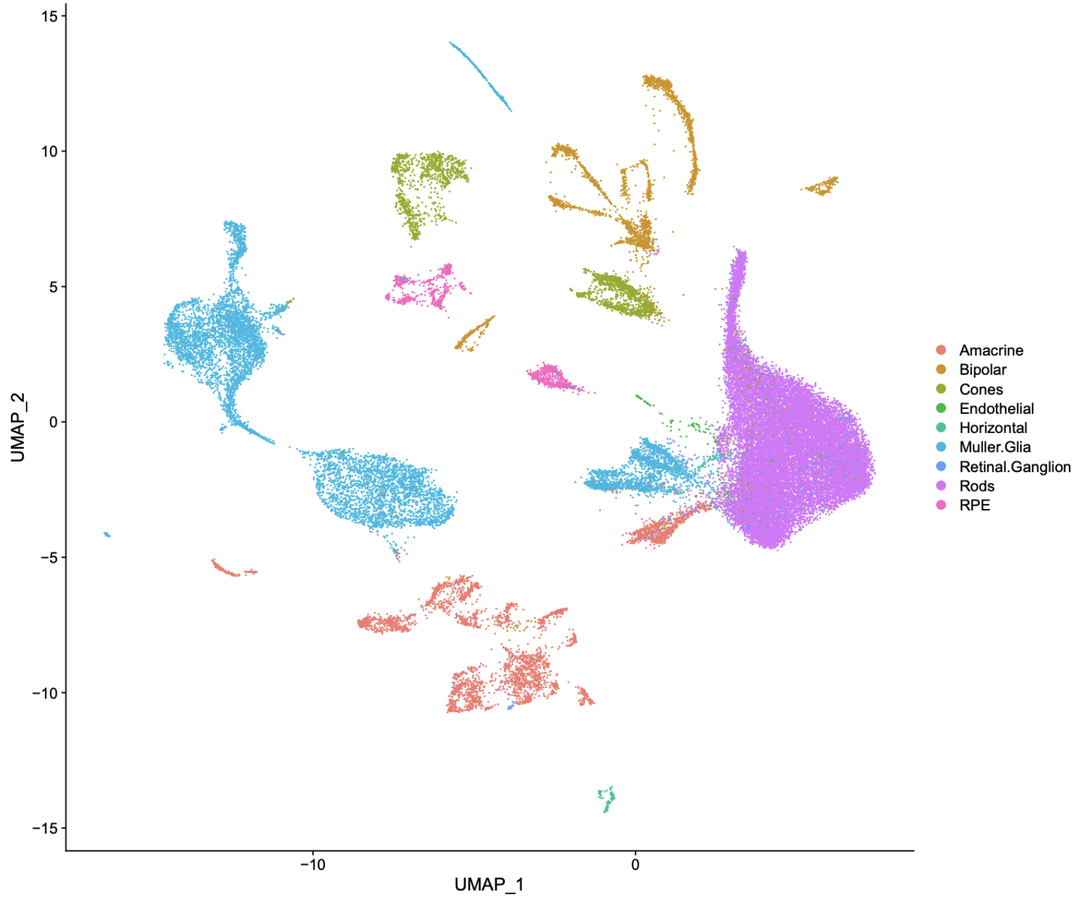


Non-linear dimensionality reduction and visualization

“Standard” processing –
LogNormalize, Scale, 3000 HVGs, 30 PCs

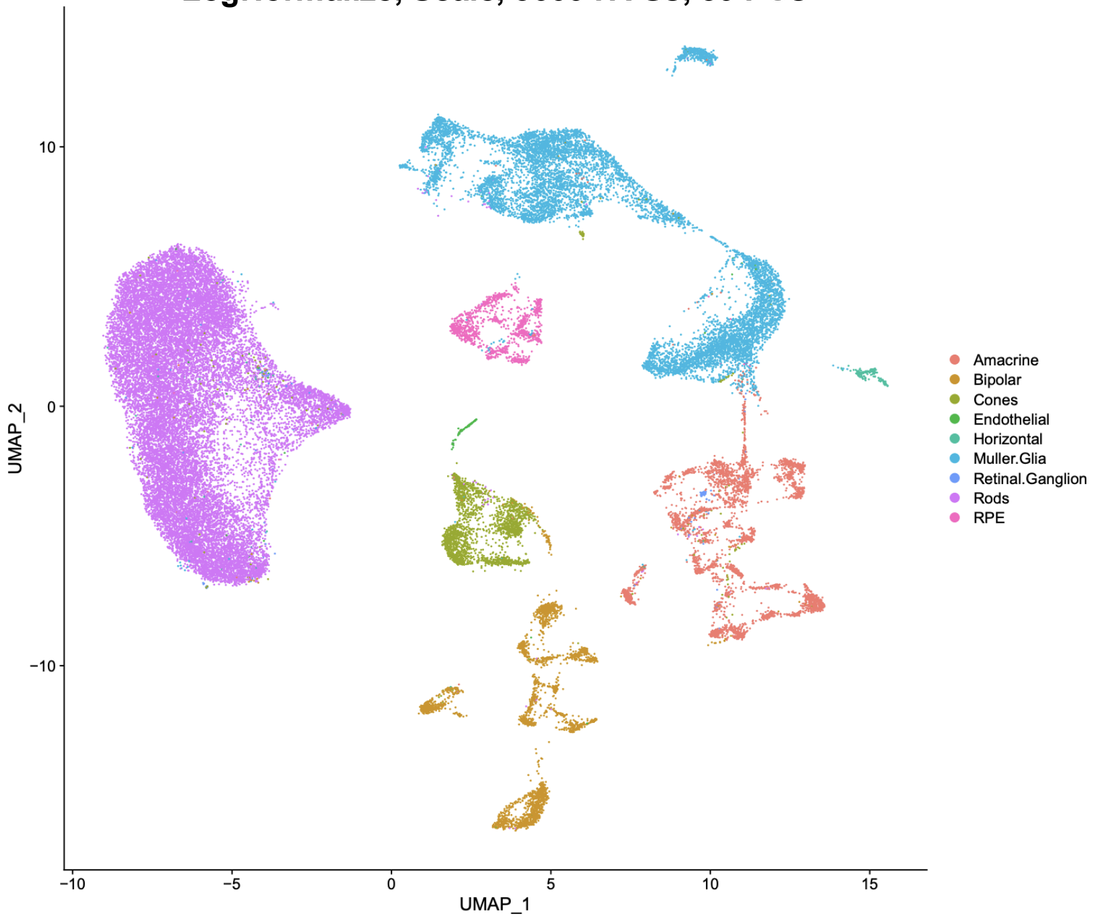


Un-normalized

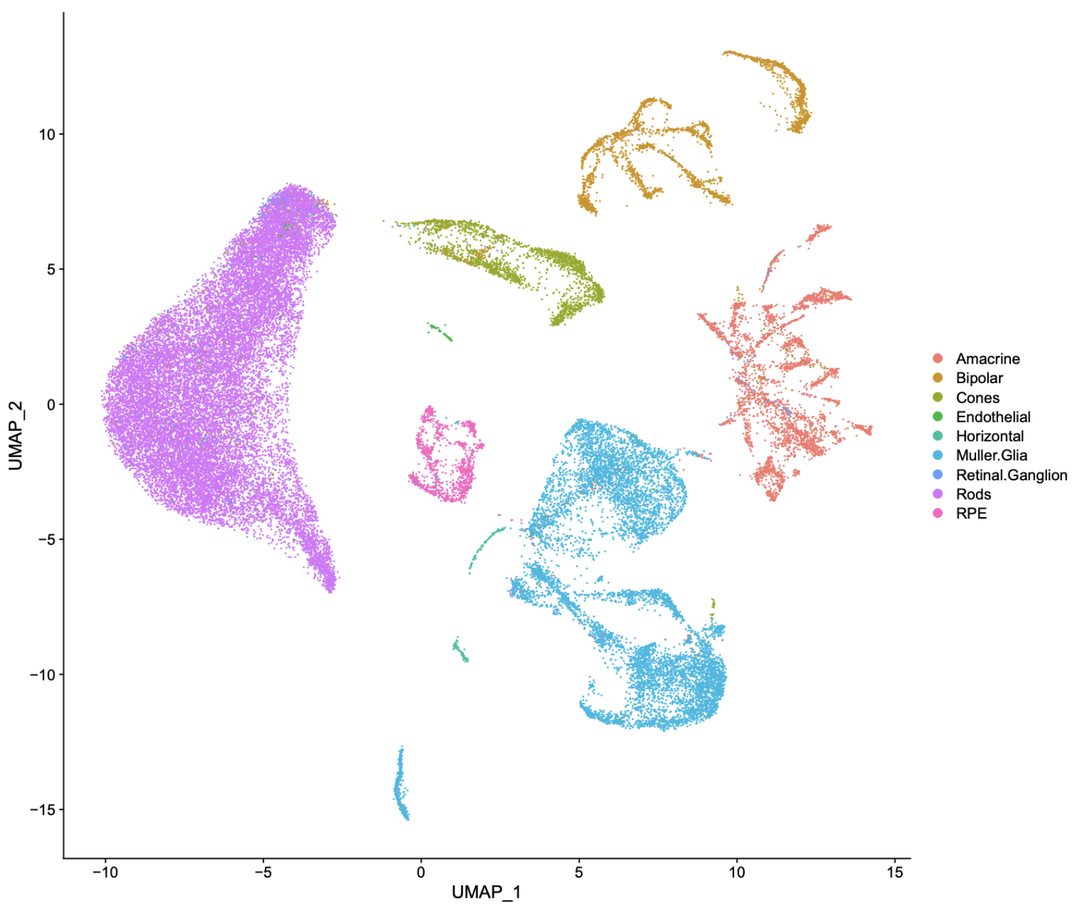


Non-linear dimensionality reduction and visualization

“Standard” processing –
LogNormalize, Scale, 3000 HVGs, 30 PCs

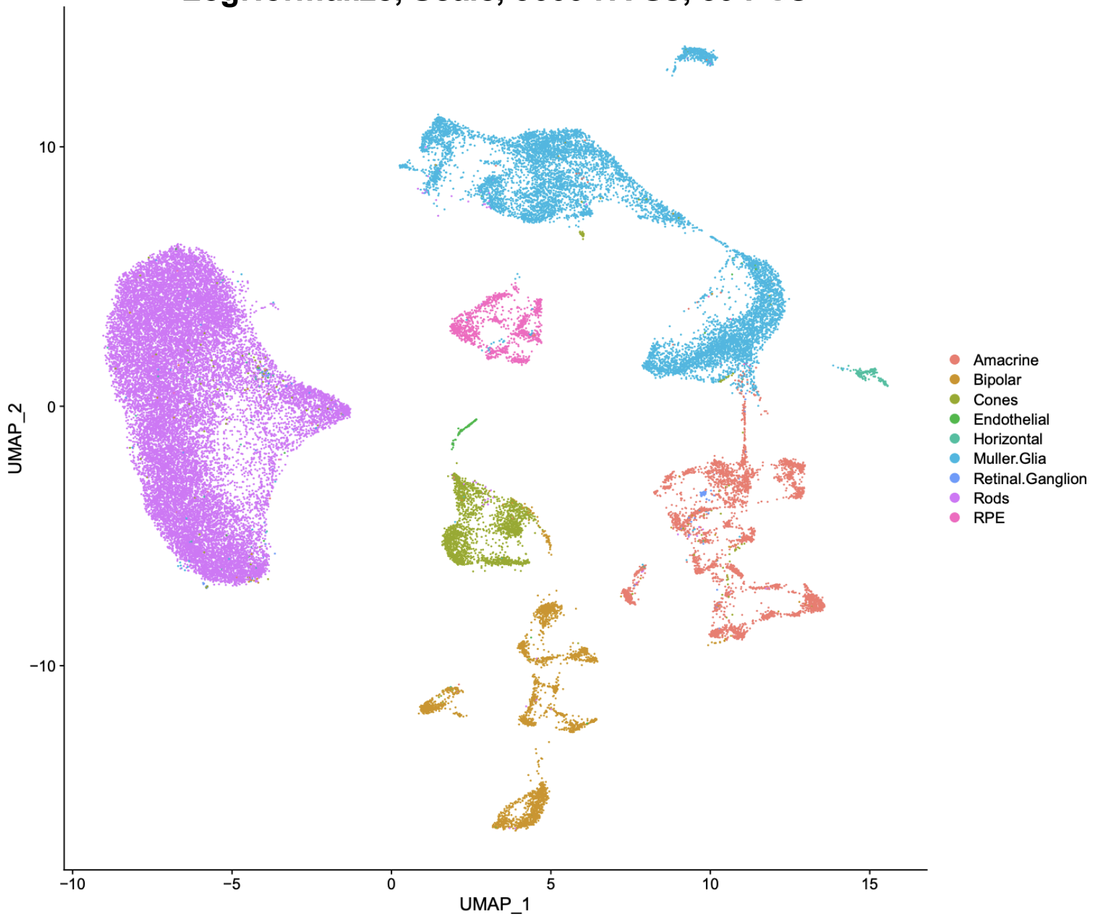


CLR Normalization

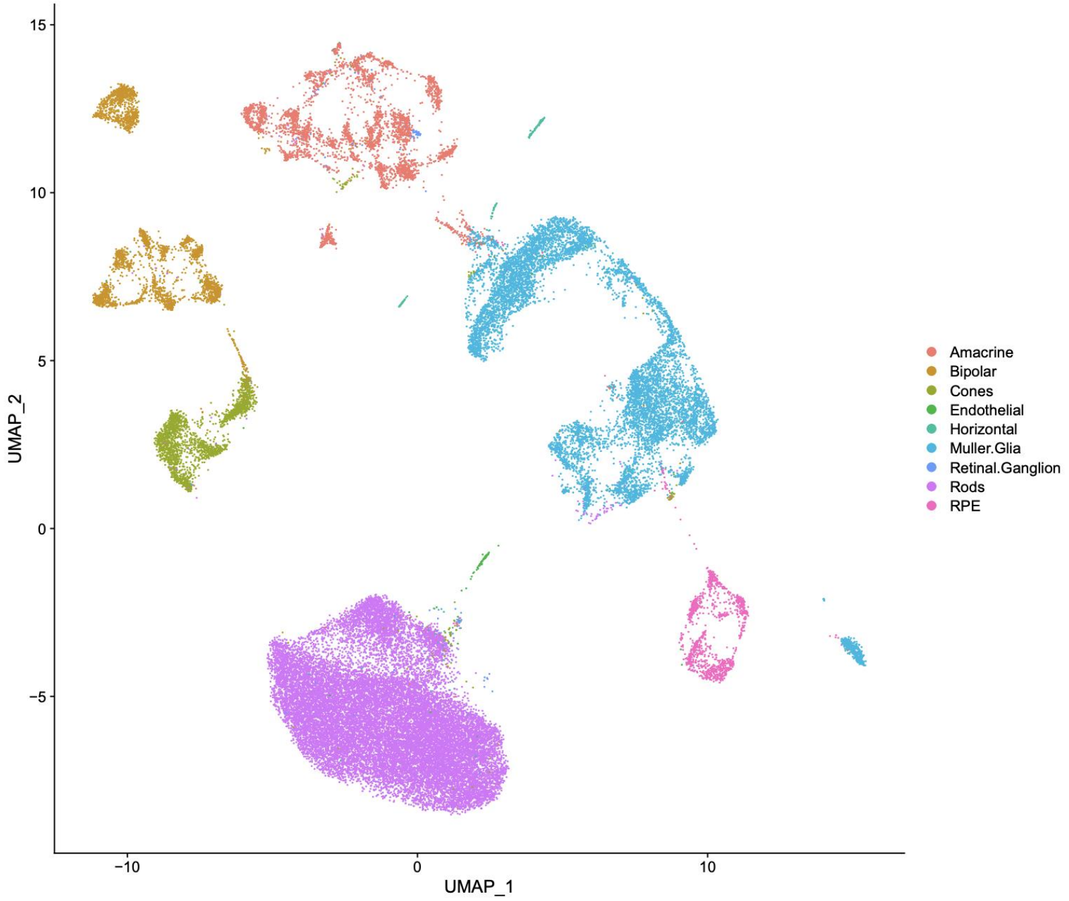


Non-linear dimensionality reduction and visualization

“Standard” processing –
LogNormalize, Scale, 3000 HVGs, 30 PCs

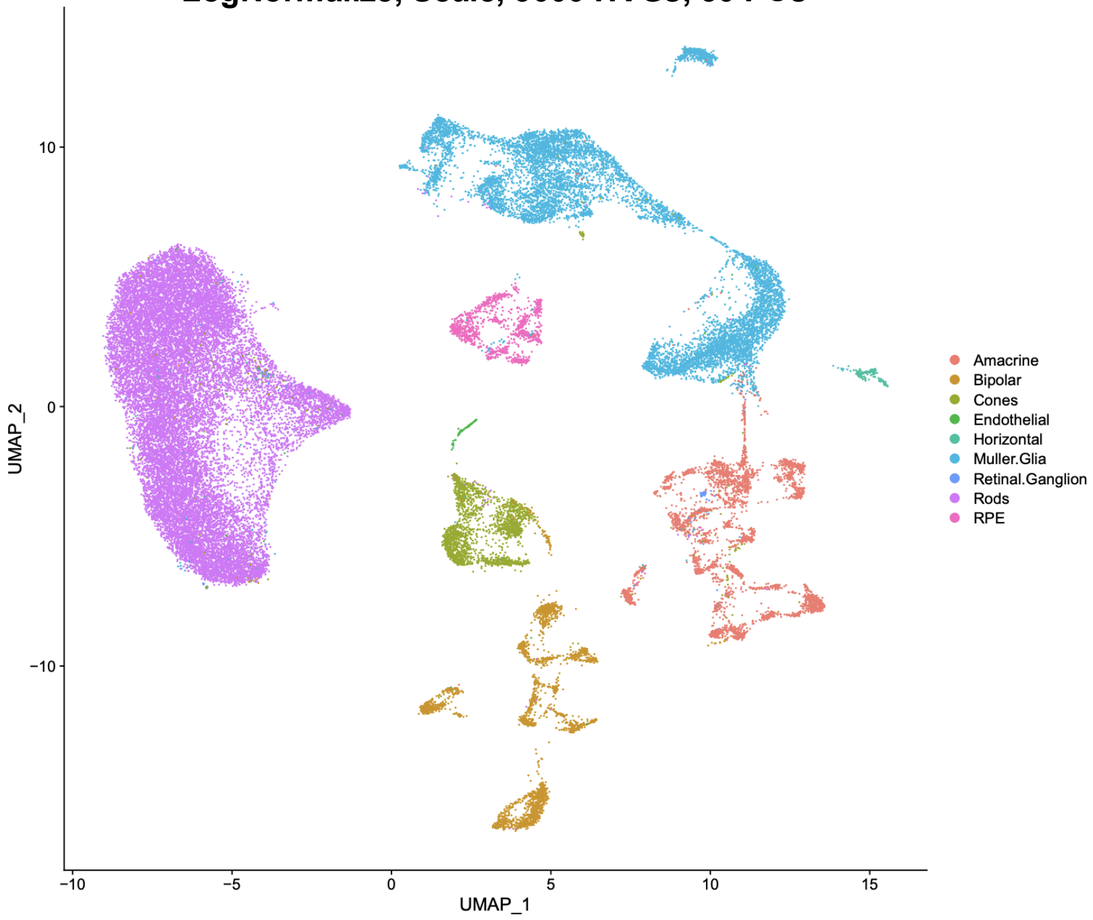


Un-scaled

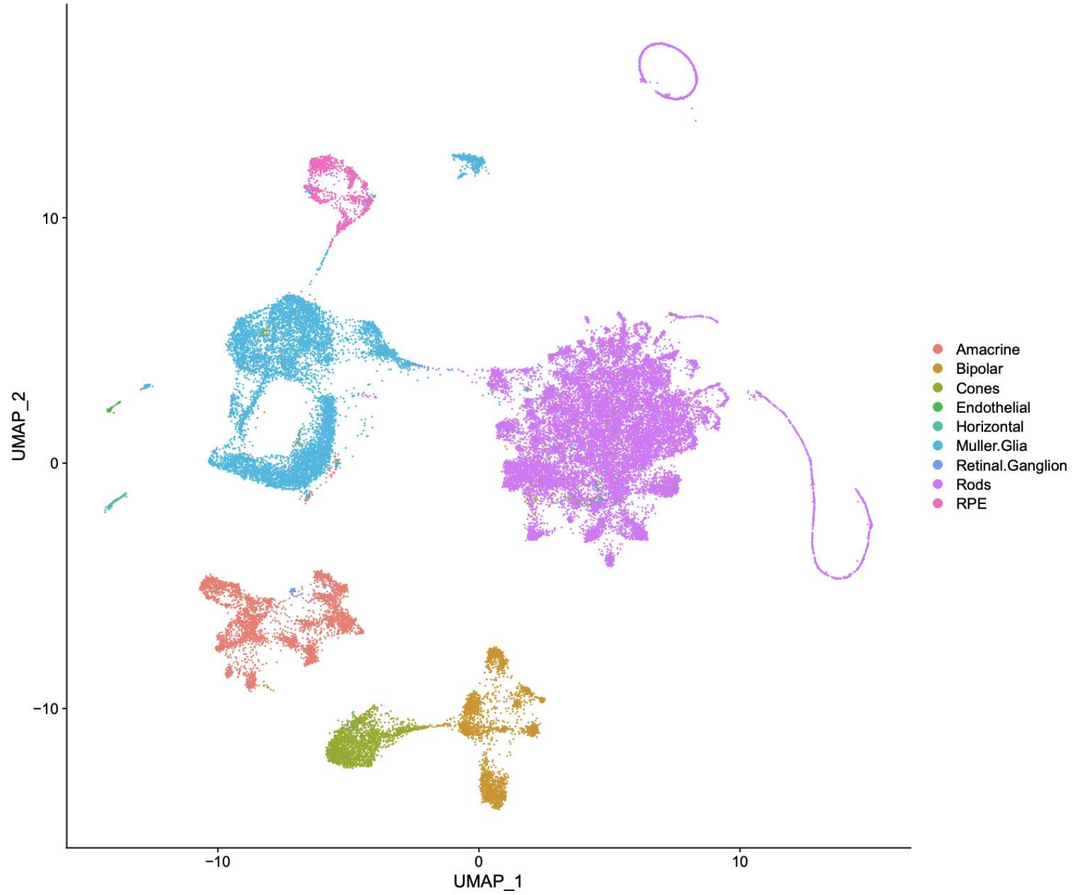


Non-linear dimensionality reduction and visualization

“Standard” processing –
LogNormalize, Scale, 3000 HVGs, 30 PCs

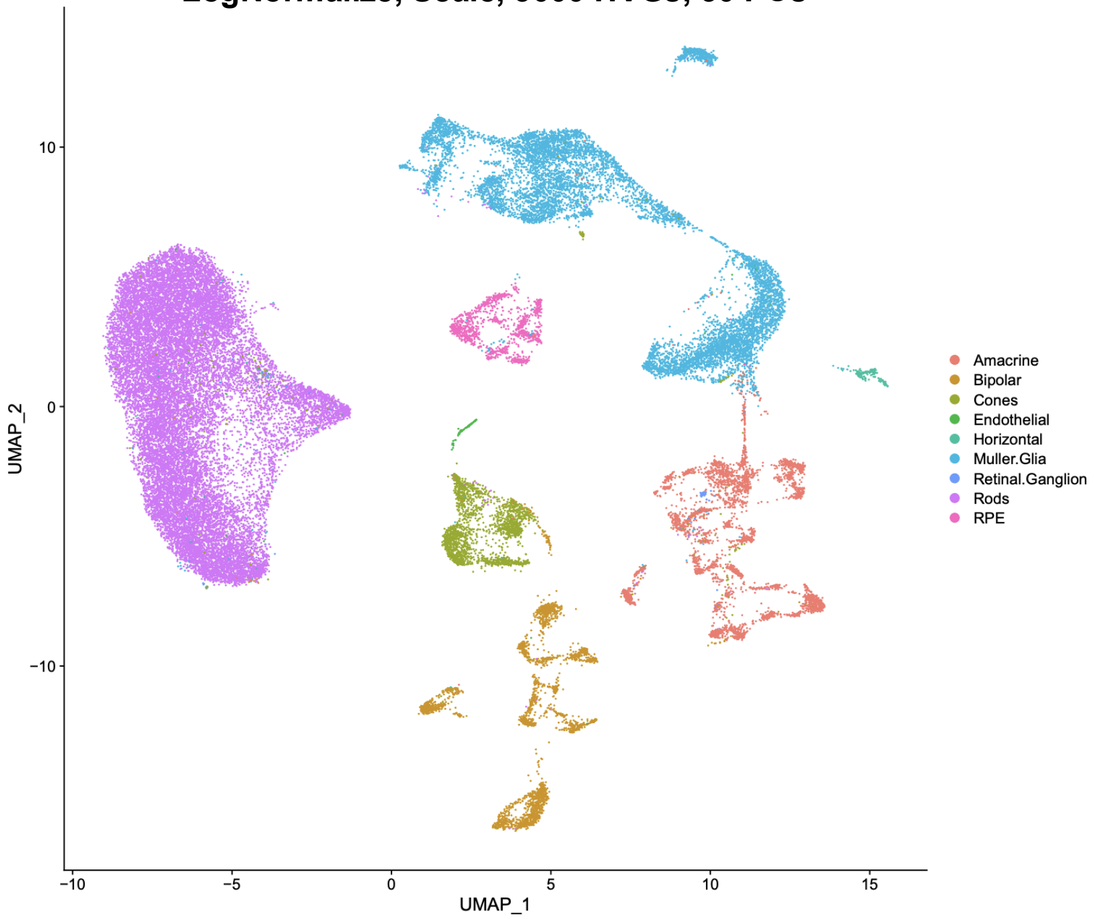


500 HVGs

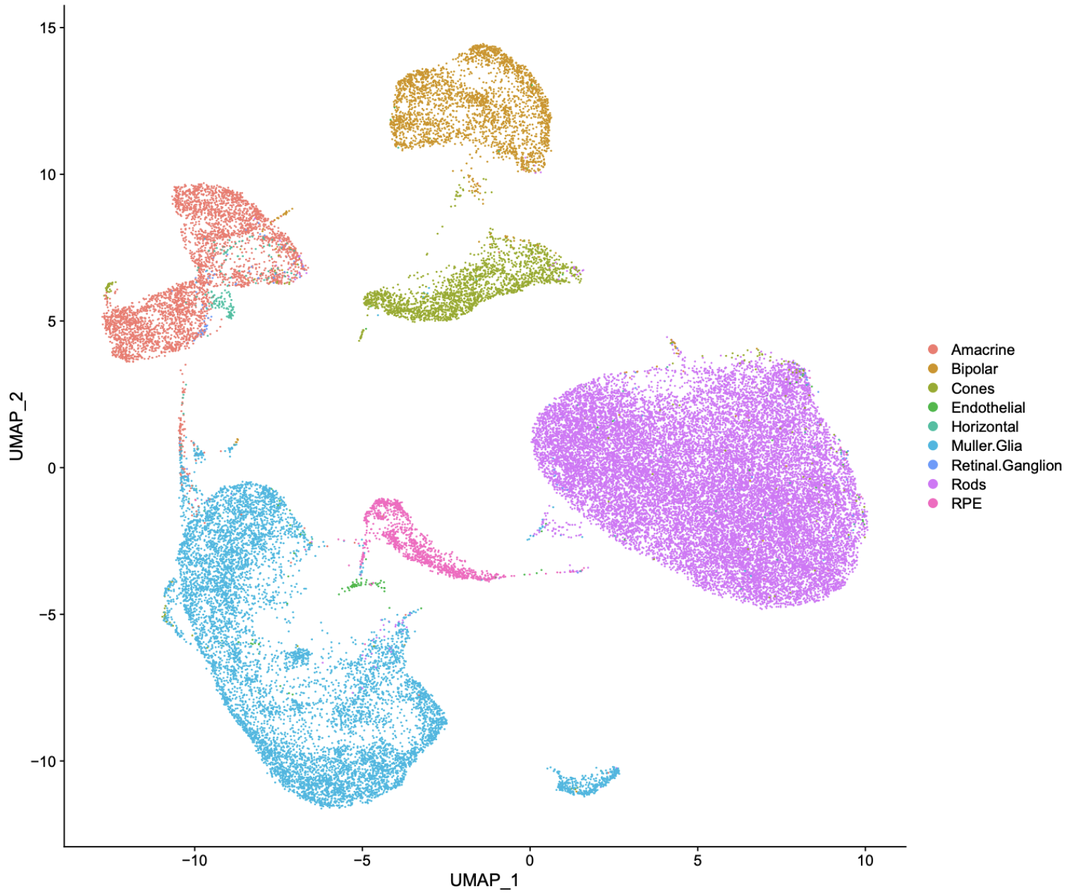


Non-linear dimensionality reduction and visualization

“Standard” processing –
LogNormalize, Scale, 3000 HVGs, 30 PCs

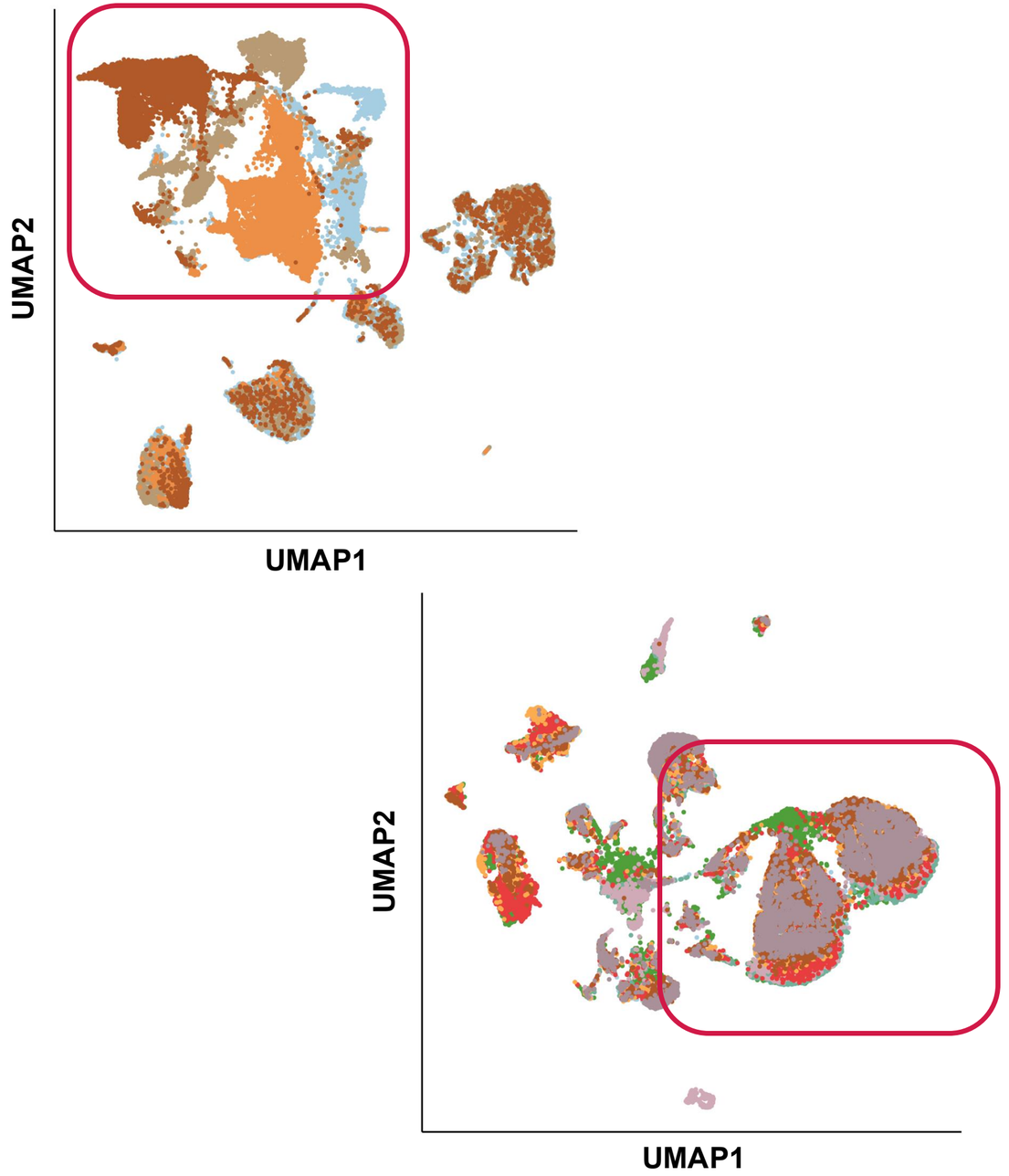


7 PCs



Integration

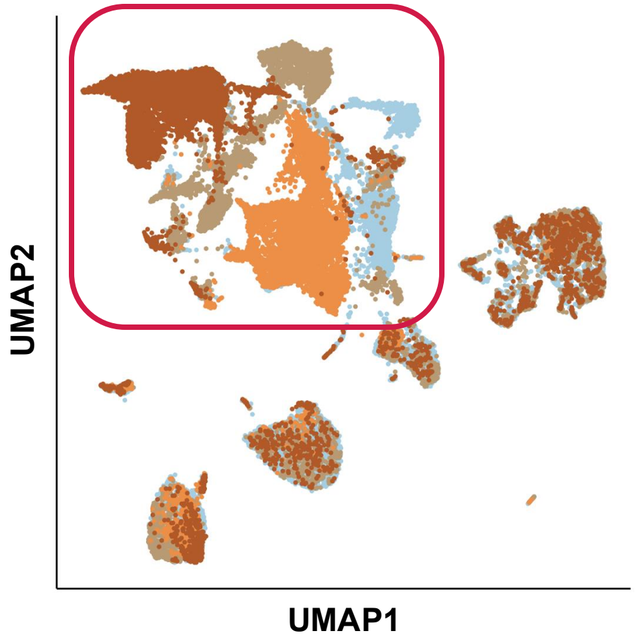
- Often group multiple samples for clustering, visualization, downstream analysis
- Simple merges show sample-based batch effects
  - Distinct cluster per sample
  - Rainbow effect within clusters

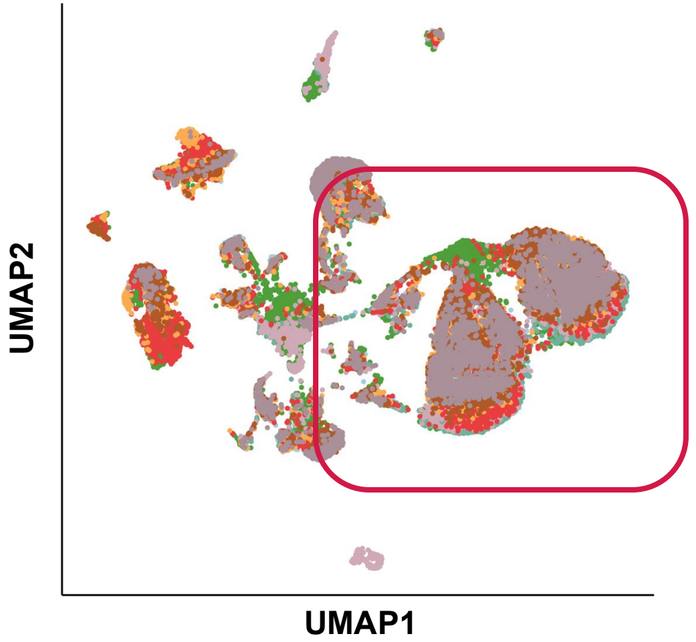


Integration

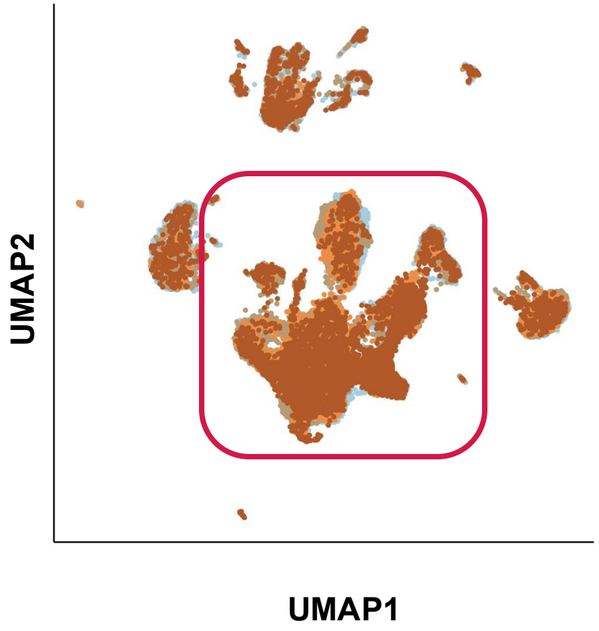
- Many tools, methods for integration
- Bioinformatics core uses three methods
  - Harmony (separate package)
  - LIGER (separate package)
  - Anchor-based integration (implemented in Seurat)
- Harmony attempts to reduce technical variation, maintain biological variation
- LIGER assumes batch effects are technical
- Step-wise anchor-based integration saves resources, can introduce bias
- No consensus on “best” method, depends on size, complexity, etc.

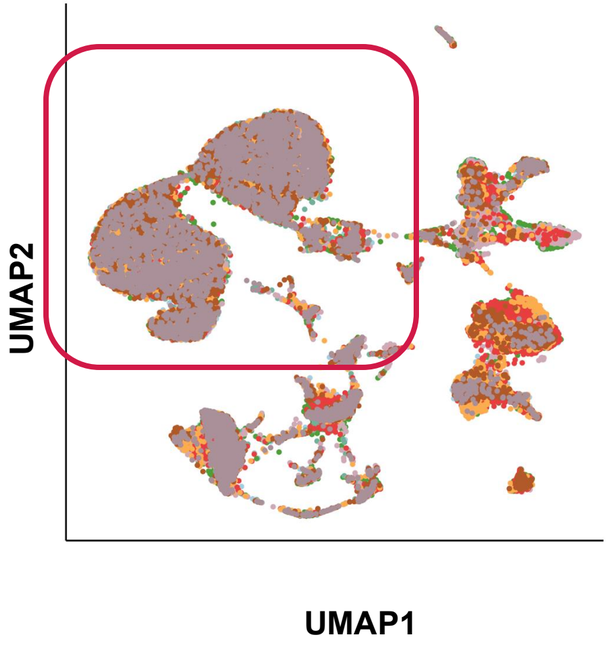
Non-integrated:





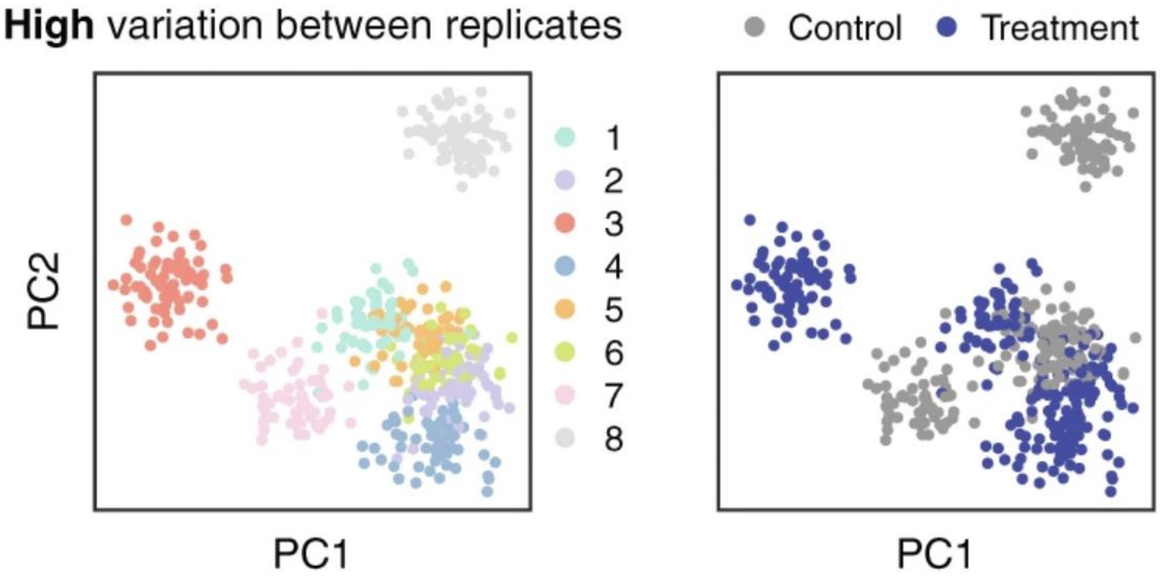
Integrated:





Differential Expression Analysis

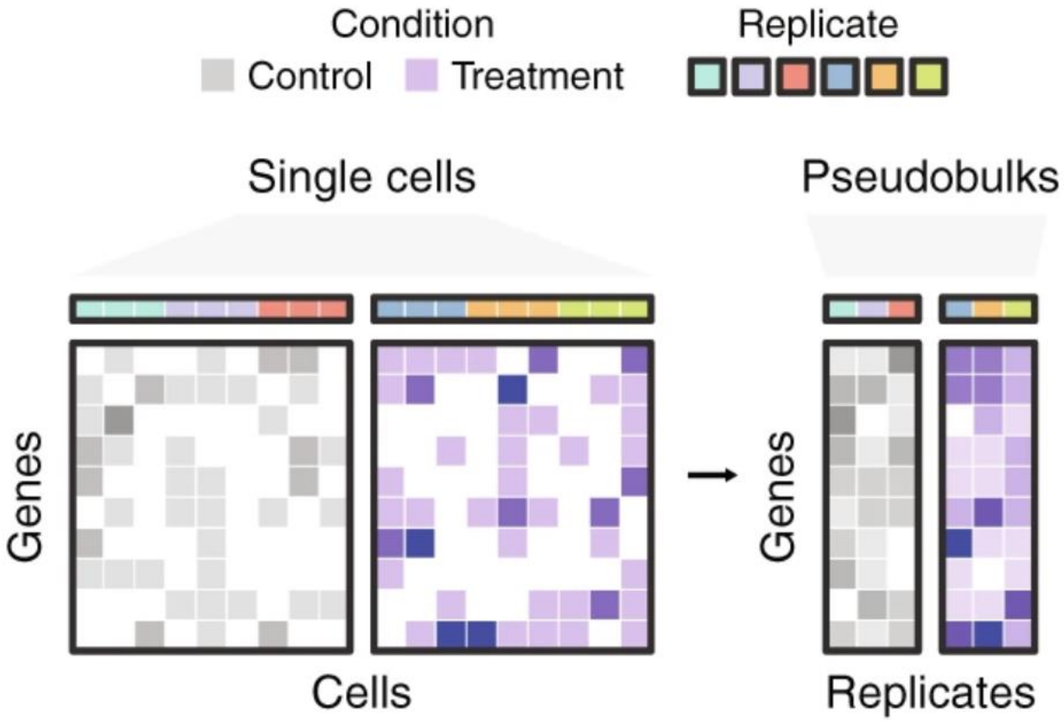
- Consider batch effects when running, interpreting DE
  - Technical and biological effects
  - Dropout, inflated zero counts
  - High variability across samples



Squair (2021). Nature Communications.

Differential Expression Analysis

- Two categories for methods:
  - Cell-level
  - Pseudobulk
- Cell-level methods treat cell expression profiles individually.
  - e.g. Wilcoxon rank sum (Seurat's default)
- Pseudobulk methods aggregate expression profiles per sample.
  - e.g. DESeq2 (bulk RNA-seq method)



Squair (2021). Nature Communications.

Differential Expression Analysis

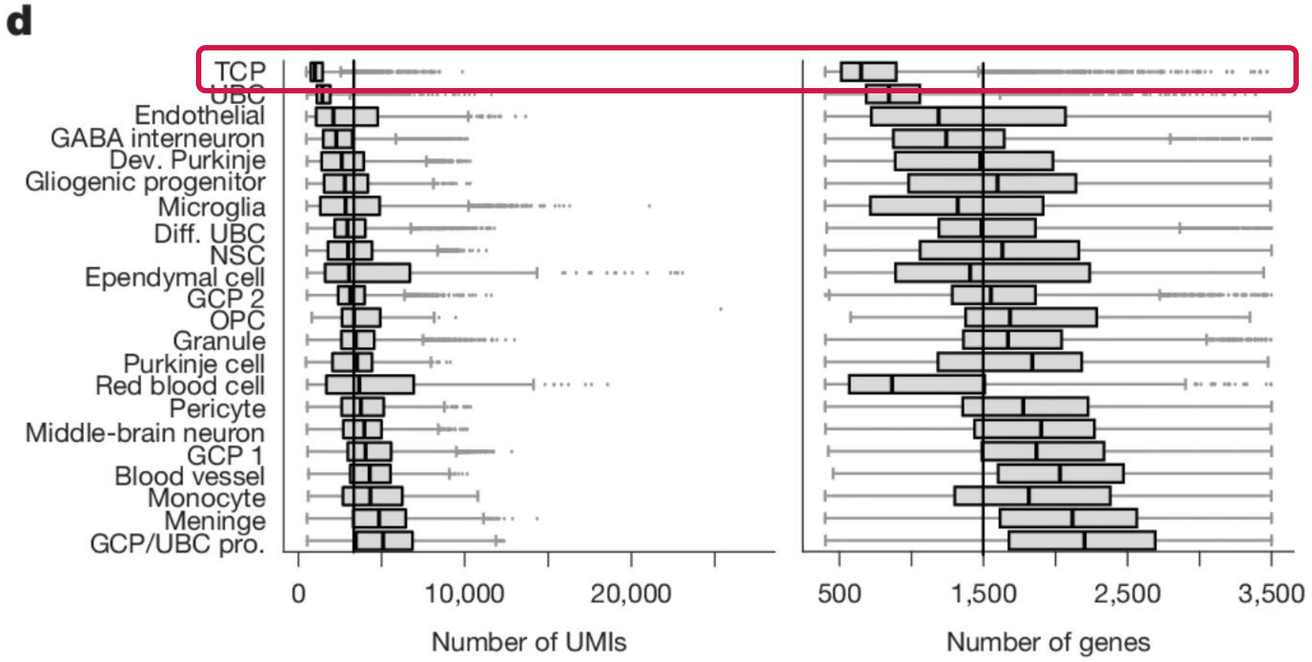
- Cell-level methods prone certain issues
  - Inflated significance of p-values
  - Inflated number of false positives
- Fail to address dependence of cells in same sample
- Still useful for broad results
  - e.g. generating per cluster markers

Differential Expression Analysis

- Pseudobulk methods address certain issues
  - Reduce inflated significance
  - Reduce false positive rates
- Allow more complex models, confounding variables
- Need at least two, ideally three replicates per variable of interest
- Can still see batch effects, inflated significance, false positives
- Reproduce results in other datasets
- Confirm via independent wet lab methods

Cell Type Annotation

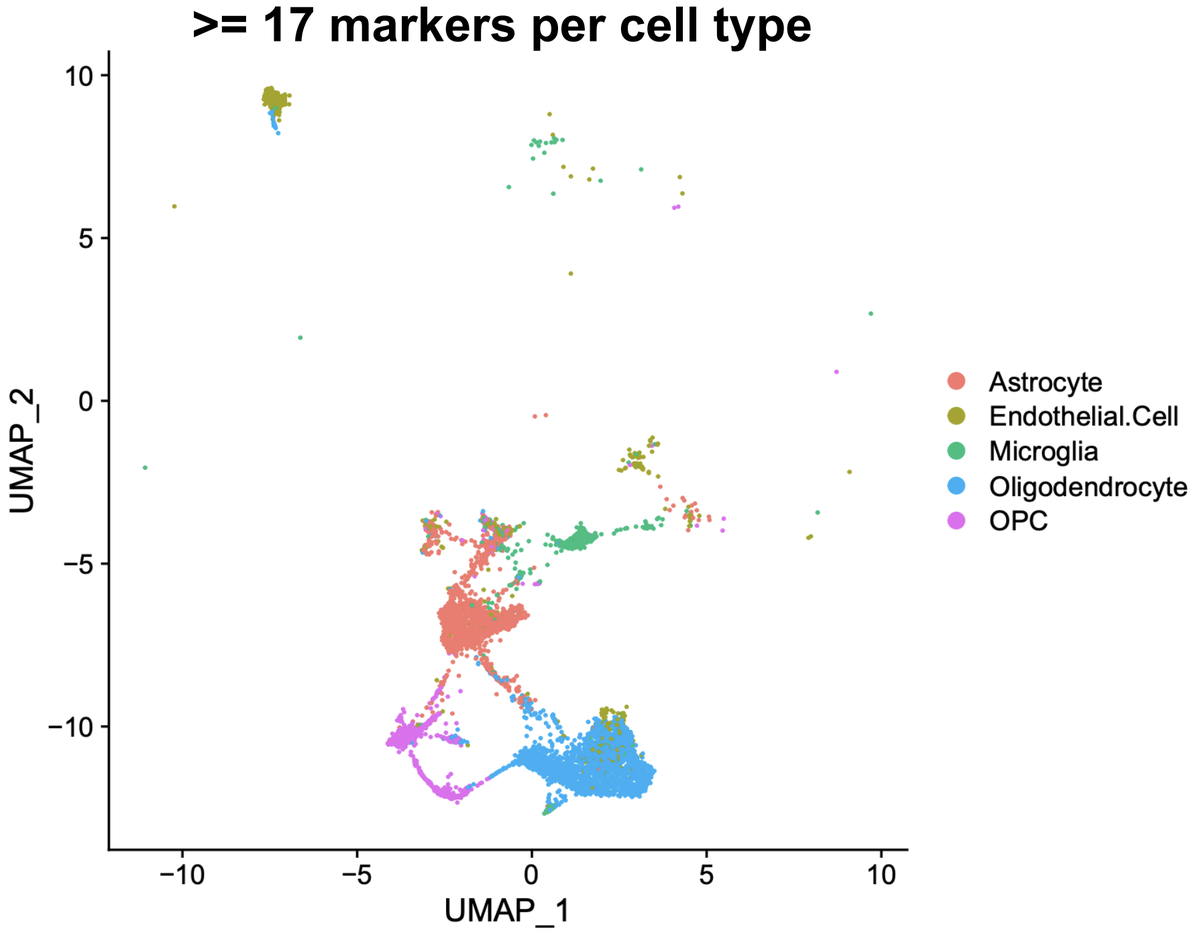
- Pitfalls upstream can lead to inaccurate annotations downstream
- Inaccurate annotations can lead to false conclusions

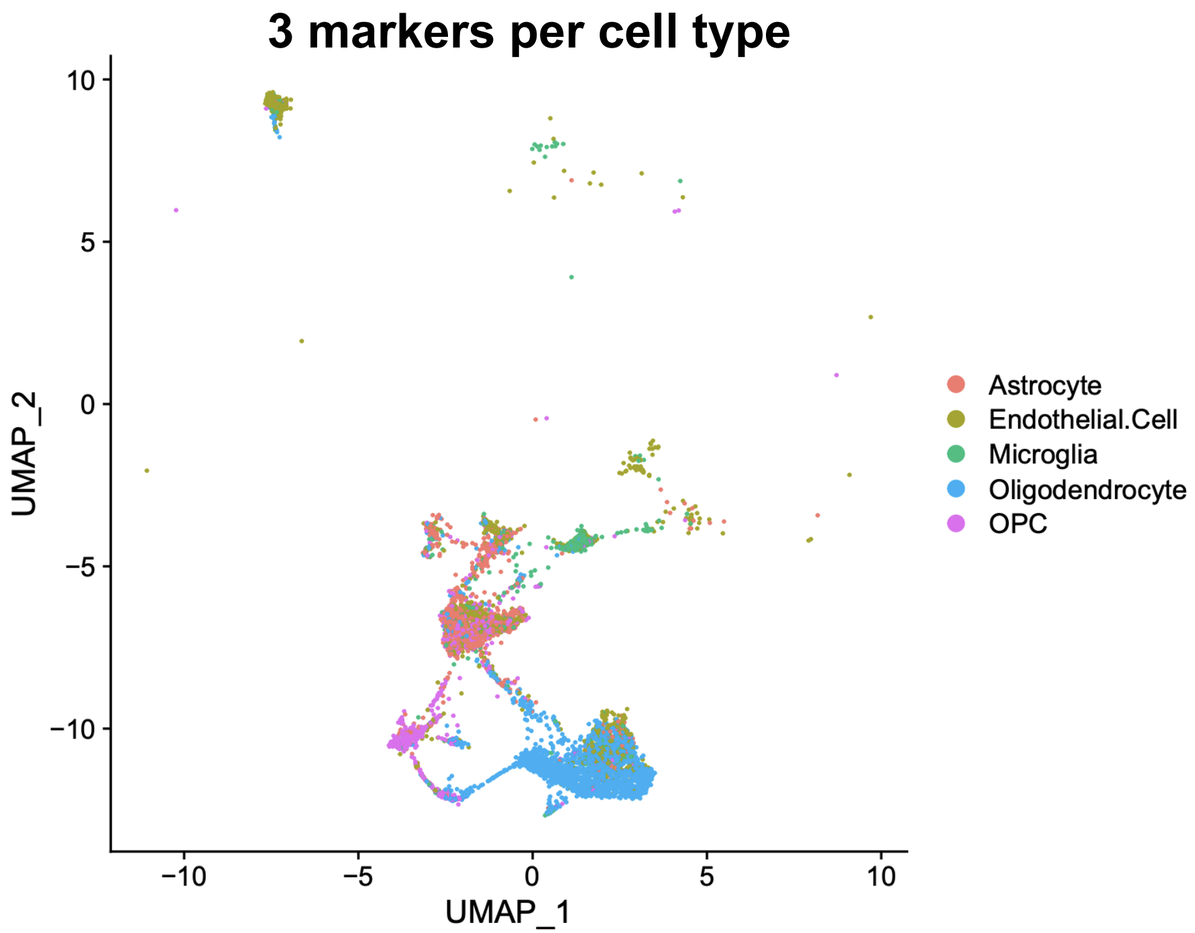


Smith (2025). Nature.

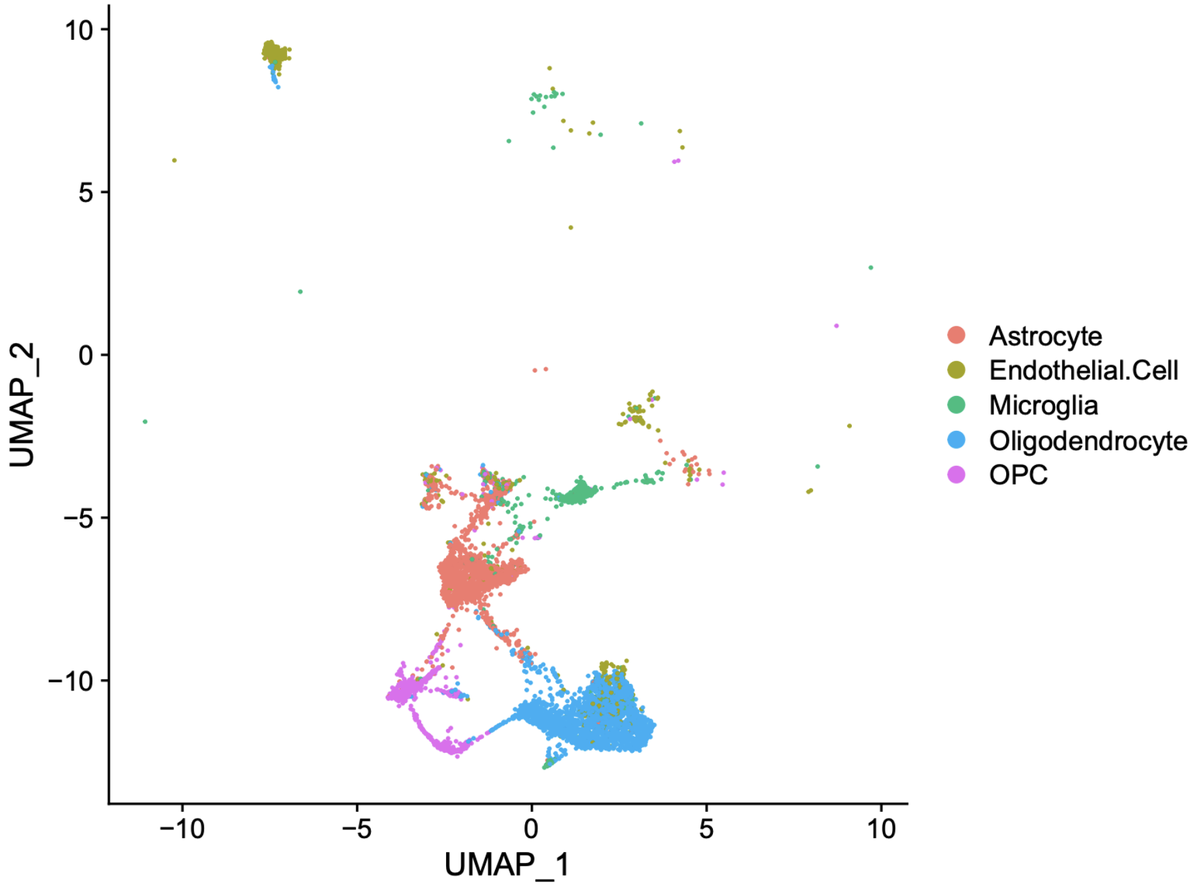
Cell Type Annotation

- Annotations can involve manual curation, semi-automated methods, fully automated methods, e.g.
  - Module scoring (semi-automated, uses curated gene marker lists)
  - SingleR annotation (automated, uses annotated single cell or bulk references, generates broad or fine labels)
  - Reference label transfer (automated, uses annotated single cell objects)
- Consider number of cell type markers use for module scoring
  - Too few, overlapping markers can cause mislabeling, miss subtypes
  - More makers reduces noise in annotations
- Module scoring will force label assignment
  - Can only assign labels provided as input

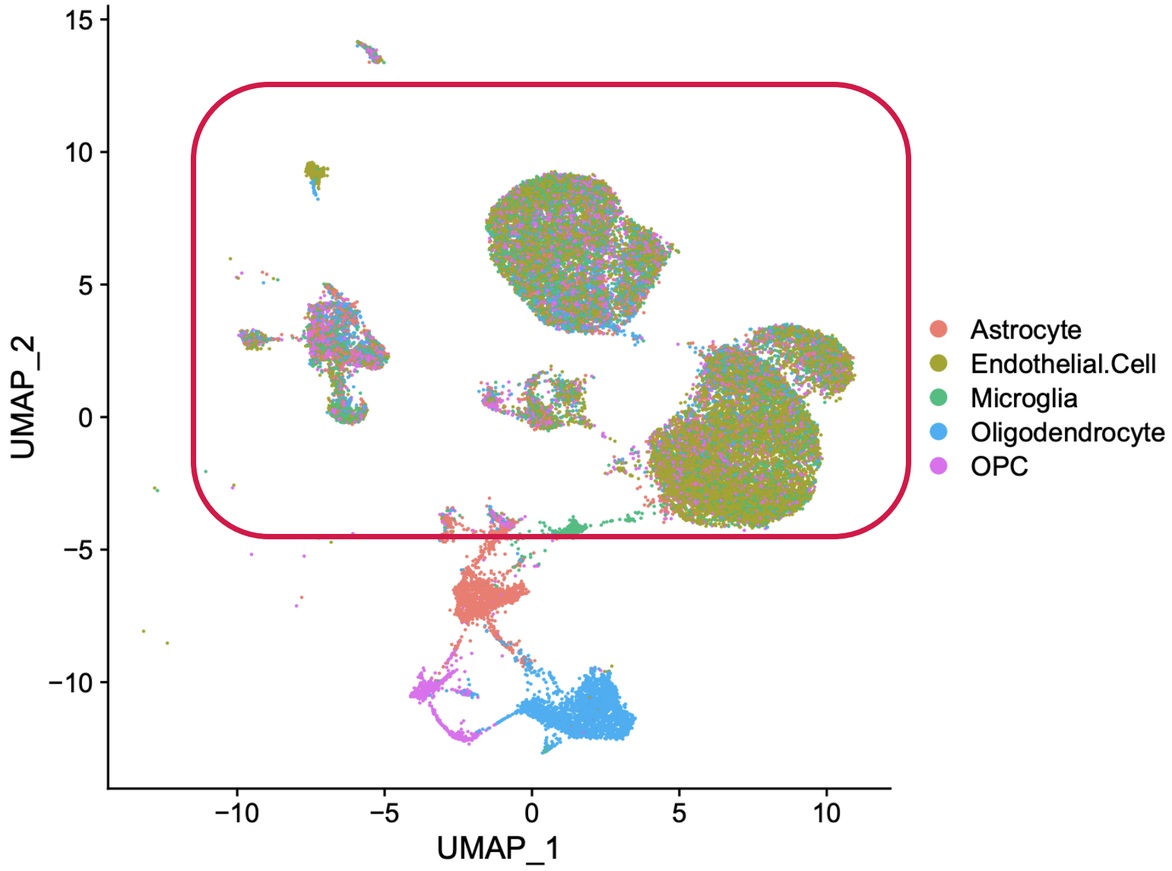

>= 17 markers per cell type
UMAP_2
UMAP_1
10
5
0
−5
−10
−10
−5
0
5
10
Astrocyte
Endothelial.Cell
Microglia
Oligodendrocyte
OPC


3 markers per cell type
UMAP_2
UMAP_1
10
5
0
−5
−10
−10
−5
0
5
10
Astrocyte
Endothelial.Cell
Microglia
Oligodendrocyte
OPC

Non-neuronal cells



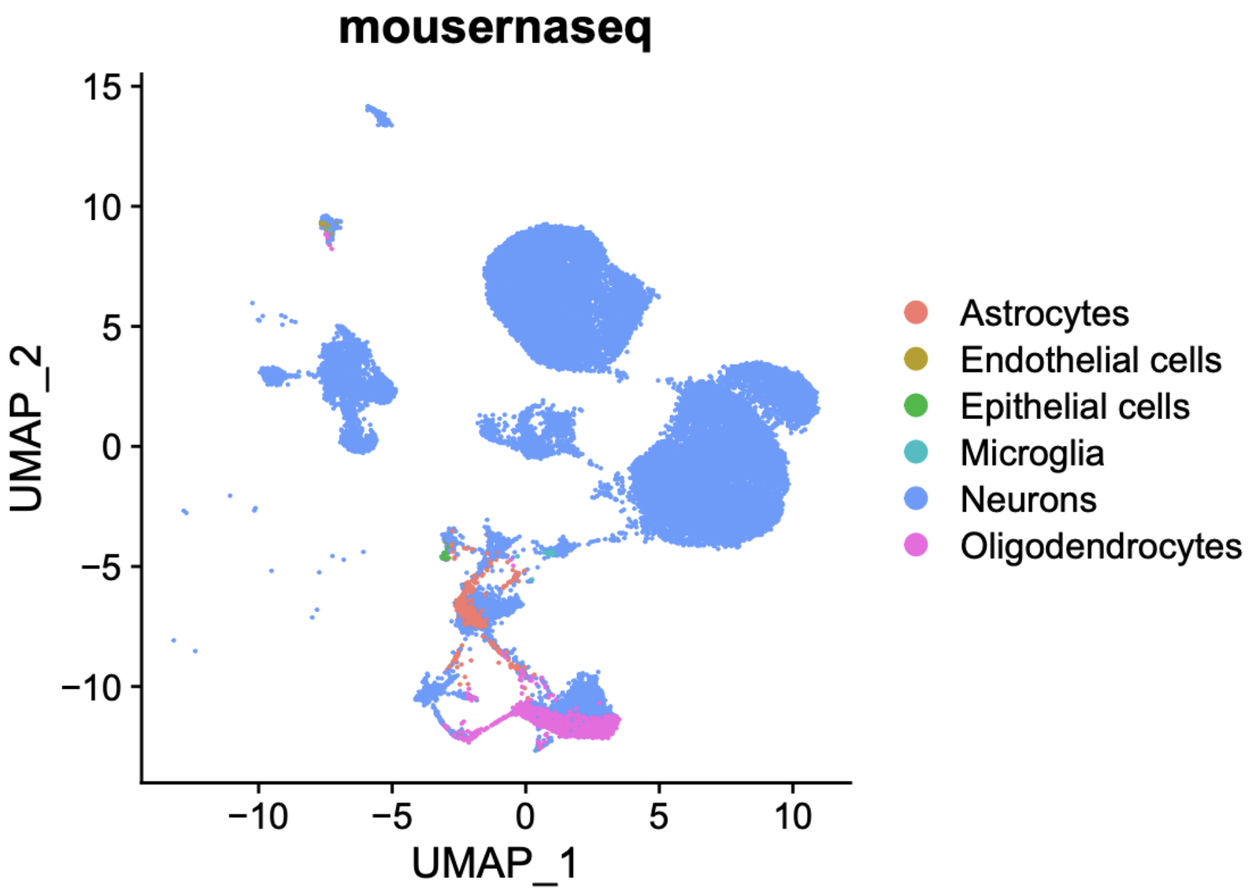
Neuronal cells

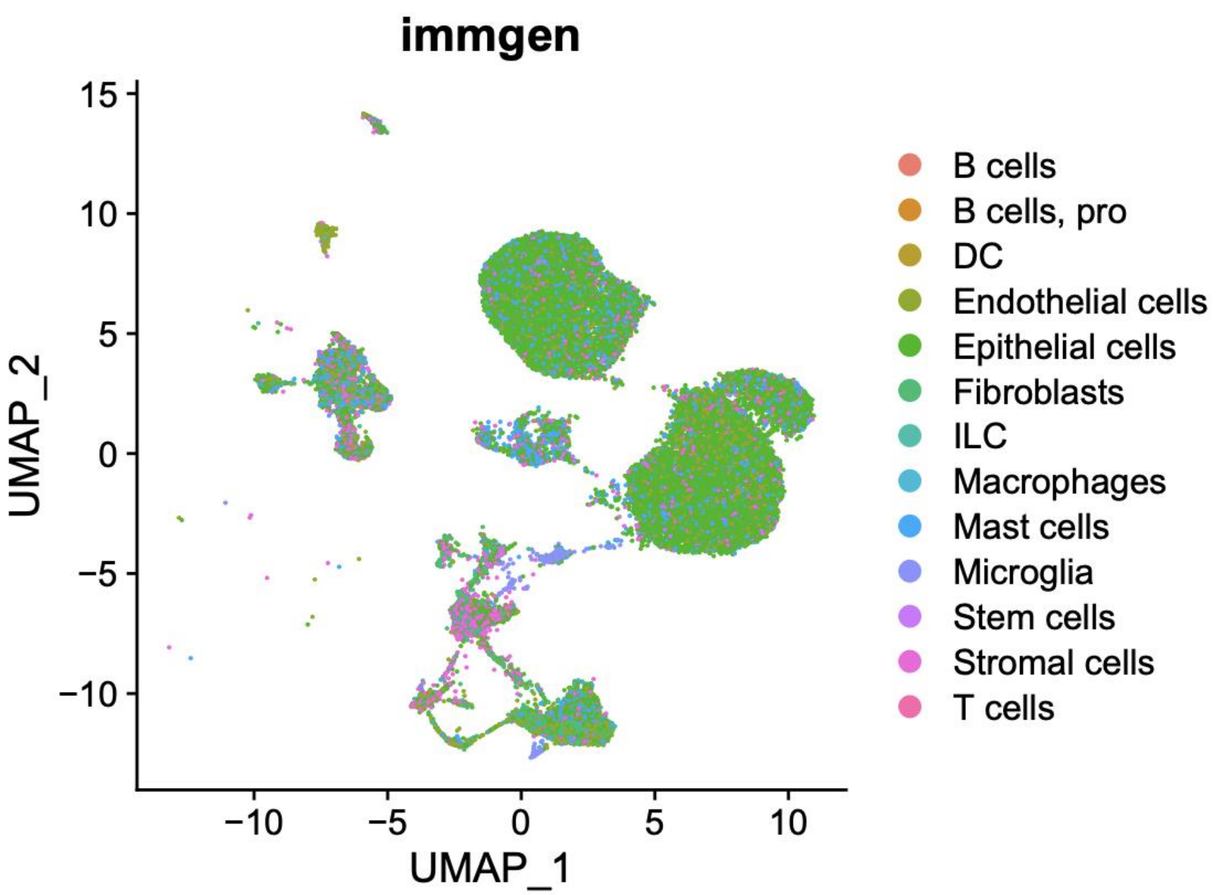


Cell Type Annotation

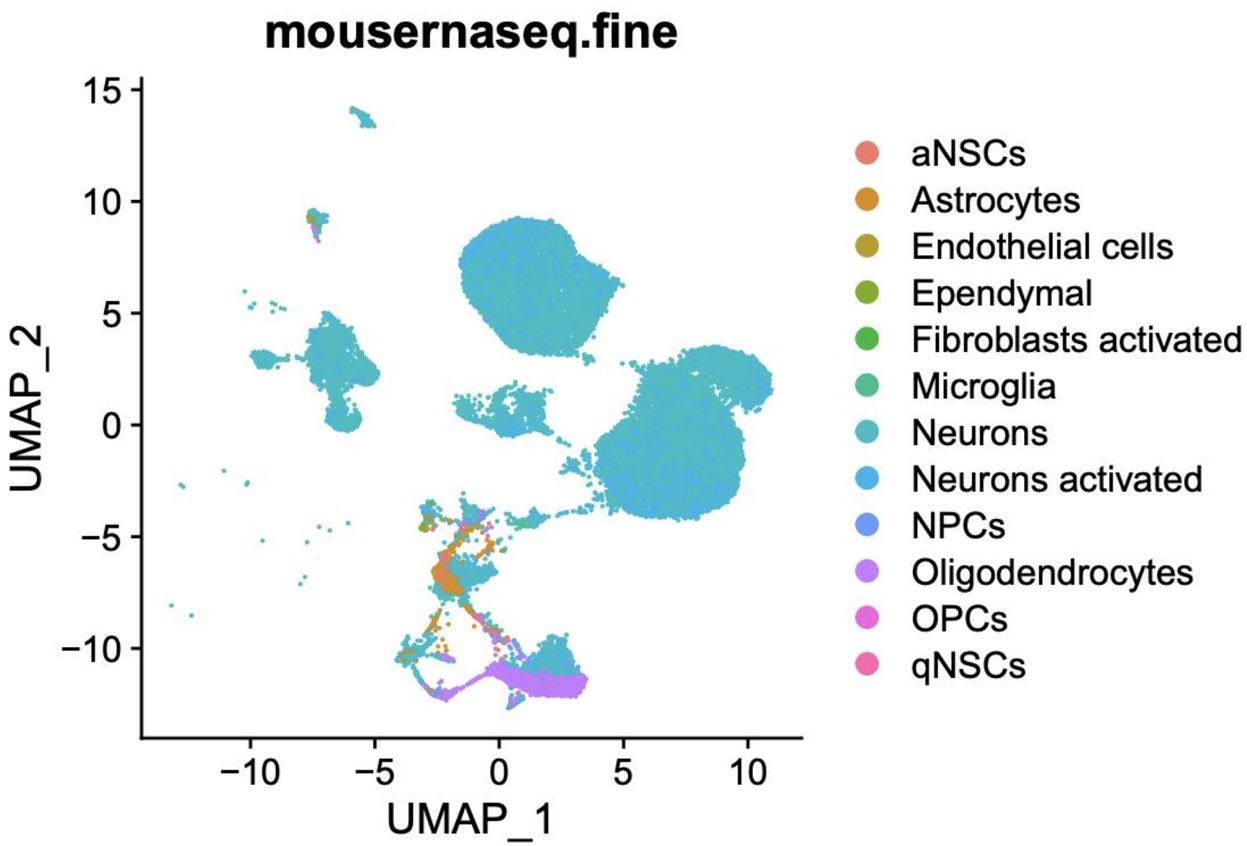
- Important to understand resources used for automated tools (e.g. SingleR, reference label transfer)
  - How are cell types annotated?
  - What cell types are present?
  - How detailed are cell type labels?
- Do methods force labels to be assigned?
- SingleR provides broad, fine labels, can prune cells
- Label transfer detail depends on reference, labels every cell
- Rely on combo of semi-automated methods, fully automated methods, manual annotation

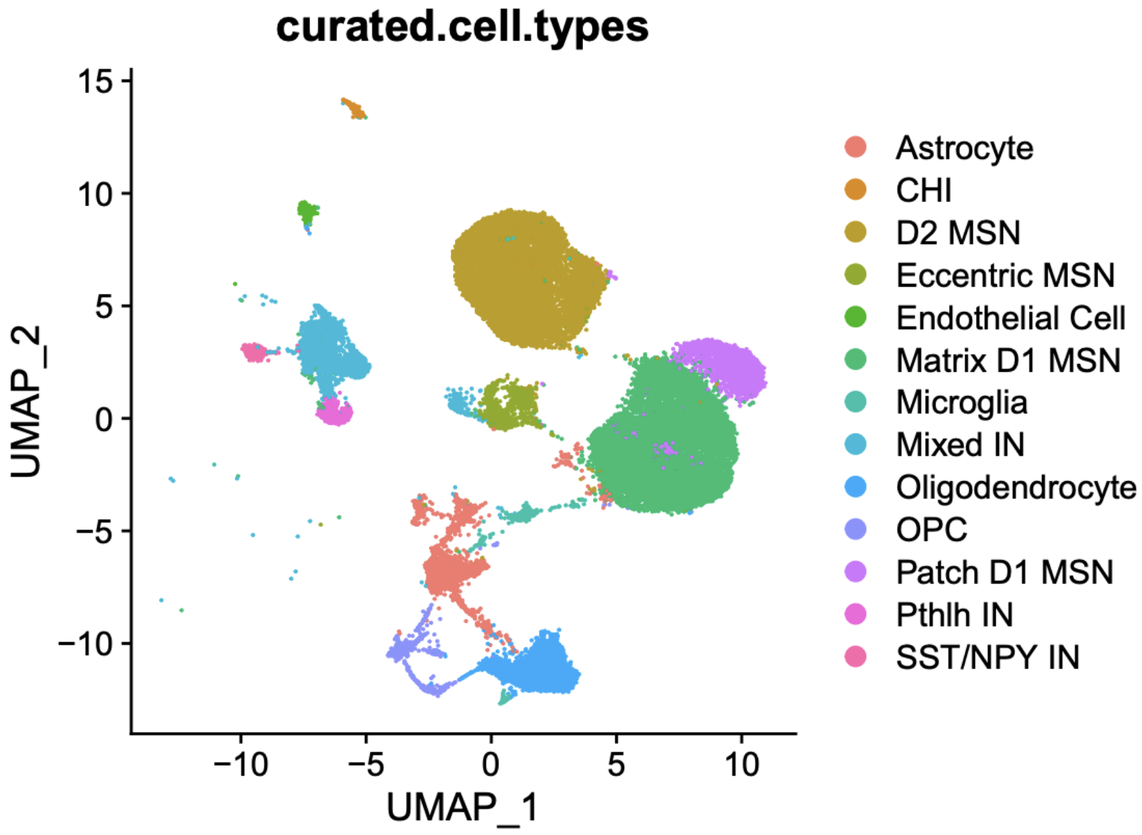
Cell Type Annotation





Cell Type Annotation





Summary

- Pitfalls can occur during most steps.
- Analysis often includes redoing, refining steps
- These pitfalls can give you a starting point for fine tuning.
- Be careful not to overinterpret data
- Be careful not to overly adjust to fit expected biology
- Verify findings with additional methods
- Gold standard is wet lab verification

Our Tools & Resources

Tool/Resource	Description	Icon
Snap	sc/snRNA-Seq 10x; supports human, mouse, and dual genomes	
Snap-10x-Flex	sc/snRNA-Seq 10x Flex; supports human and mouse genomes	
sc-epigenie	scATAC-Seq 10x Flex; supports human and mouse genomes	
sc-atac-seq-bed-builder	Generate blacklist/promoter/enhancer BED files for analysis	
DevOps Containers	Reusable, independent container images for scRNA/ATAC-seq workflows	
Trainings & Workshops	Hands-on training courses and pipeline tutorials